Astronomy Reports, Vol. 47, No. 12, 2003, pp. 1000–1012. Translated from Astronomicheskiĭ Zhurnal, Vol. 80, No. 12, 2003, pp. 1086–1098.
Original Russian Text Copyright © 2003 by Efremov.

Cepheids in LMC Clusters and the Period–Age Relation

Yu. N. Efremov

Sternberg Astronomical Institute, Universitetskiĭ pr. 13, Moscow, 119992 Russia

Received May 28, 2003; in final form, July 26, 2003

Abstract—We have made a new comparison of the positions of Cepheids and clusters in the LMC and constructed a new empirical period–age relation taking into account all available data on Cepheids in the LMC bar provided by the OGLE project. The most probable relation is $\log T = 8.50 - 0.65 \log P$, in reasonably good agreement with theoretical expectations. Numerous Cepheids in rich clusters of the LMC provide the best data for comparing theories of stellar evolution and pulsation and the dynamical evolution of clusters with observations. These data suggest that stars undergoing their first crossing of the instability strip are first-overtone pulsators, though the converse is true of only a small fraction of first-overtone stars. Several rich clusters with suitable ages have no Cepheids—a fact that is not understood and requires verification. Differences in the concentration of Cepheids toward their cluster centers probably reflect the fact that the clusters are at different stages of their dynamical evolution, with the Cepheids in cluster coronas being ejected from the cluster cores during dynamical interactions between stars.
© 2003 MAIK "Nauka/Interperiodica".

1. INTRODUCTION

The discovery of Cepheids in open clusters of our Galaxy was a very important milestone in recognizing the nature of these stars, and made it possible to solve the problem of the zero point of the period–luminosity relation and provide a basis for the hypothesis that Cepheids originate from main-sequence B stars. This hypothesis predicts that there should exist a certain relation between the periods of Cepheids and the ages of clusters containing them. According to stellar-evolution theory, stars of higher mass leave the main sequence more rapidly and, due to their lower density, pulsate with longer periods when in the instability strip. Comparisons of the ages of several Galactic clusters and the periods of their Cepheids led to the first observational support for such a relation [1].

Much more certain results were obtained later, taking into account numerous Cepheids in rich clusters of the Magellanic Clouds. Some LMC clusters contain up to a dozen or more Cepheids (compared to the no more than three Cepheids observed in Galactic clusters), making it possible to draw conclusions about the real ("cosmic") dispersion of the period–age relation, since the scatter in the ages of cluster stars of similar masses is small [2].

The amount of data available on LMC clusters and their Cepheids has greatly increased in recent years. Determinations of integrated color indices that can be used to derive ages now exist for 624 clusters and associations. Searches for Cepheids have been conducted in several rich clusters. In particular, the results of the OGLE project, which have been published in detail, have led to nearly exhaustive searches for Cepheids (97% complete) and stellar clusters in the LMC bar (as a byproduct of the main research goal). This enables, for the first time, studies of the spatial relations between Cepheids and clusters of the same age, with no bias due to observational selection effects.

Cepheid periods are the best age indicators for single stars, and can be used to reconstruct the history and character of star formation in the region of the galaxy in which they are located. For example, Cepheids were used to establish the existence of large stellar complexes in our Galaxy, and of an age gradient across a spiral arm in the Andromeda galaxy. OGLE data have recently shown the existence of two regions in the eastern part of the LMC bar, each with diameters of about 300 pc, one containing many Cepheids and many clusters (with their Cepheids having the same age as Cepheids outside clusters), and the other possessing a very high density of Cepheids but only a few poor clusters. In the second case, we may be dealing with the isolated formation of massive stars [3].

Better knowledge of Cepheid ages and the duration of the Cepheid stage are required if such analyses are to reach their potential. The time is now ripe for a new comparison of data for LMC Cepheids and clusters and the derivation of a new calibration of the period–age relation. Apart from the appearance of new observational data, this is now possible due to progress in the development of stellar-evolution

1063-7729/03/4712-1000$24.00 © 2003 MAIK "Nauka/Interperiodica"

theory and the appearance of new, improved methods for the determination of cluster ages from integrated color indices. In the current paper, we present a new comparison of the spatial distributions of Cepheids and stellar clusters in the LMC and derive the parameters of the period–age relation from the Cepheids in these clusters. We also discuss some questions concerning the evolutionary stage of Cs Cepheids and the dynamical evolution of clusters, based on the presence of numerous Cepheids in LMC clusters and their coronas.

2. NEW DATA ON LMC CEPHEIDS AND CLUSTERS

It has recently become possible to determine the ages for 600 fairly rich LMC clusters on a uniform scale, based on the integrated UBV photometry of [4]. An improved technique for determining cluster ages from their integrated color indices was developed in [5]. As in our paper [2], this technique uses a combination of the $U-B$ and $B-V$ color indices, calibrated using the colors of rich clusters with known color–magnitude diagrams, enabling the derivation of ages from theoretical isochrones.

Mateo *et al.* [6] undertook systematic searches for Cepheids in a number of rich LMC clusters. In particular, they found 14 Cepheids in NGC 2031 and 8 Cepheids in NGC 2136, but the complete results of their study have not been published [7]. An important outcome of this study is that it established the absence of Cepheids in the rich clusters NGC 1711, NGC 2025, and NGC 2041. This result was unexpected, since these clusters' ages and integrated luminosities (and hence their masses) are within the range of ages and luminosities of clusters having numerous Cepheids (Fig. 1).

It is extremely important to understand the origin of this anomaly. It is not ruled out that these clusters are too dense and the amplitudes of their Cepheids too low for their variability to be detected. Another possible explanation is differences in the chemical abundances of clusters with similar ages. In the case of clusters whose ages correspond to the shortest Cepheid periods, the loops of the evolutionary tracks for stars with high metallicities are displaced to the right, and may not penetrate the instability strip. The first crossing of the instability strip is very short, making the probability to observing a star during this crossing very low.

The number of Cepheids that are possible members of LMC clusters has increased considerably thanks to the results of the OGLE project, which is aimed at searches for microlensing effects in the LMC bar. Pietrzynski and Udalski [8] have published a list of 204 Cepheids located in clusters (with distances

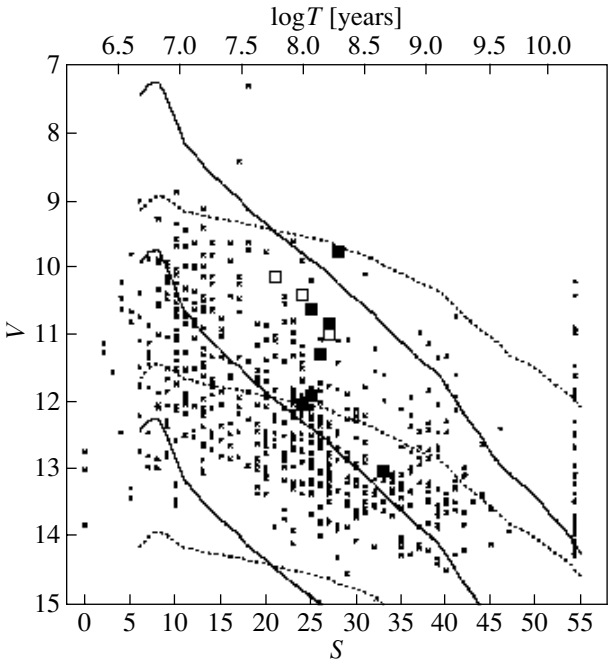

Fig. 1. Age–luminosity diagram for LMC clusters from [5], with rich clusters containing many Cepheids (filled squares) and clusters containing no Cepheids (open squares) indicated. V magnitudes are plotted along the vertical axis, while the parameter S (defined as a combination of color indices [5]; bottom) and the cluster ages ($\log T$; top) are plotted along the horizontal axis.

from the cluster centers not exceeding 1.5 cluster radii). Some of these were discovered independently in the EROS project [9]. In the LMC bar, 745 clusters were found in the field covered by the OGLE project, 126 of them new; ages have been determined for ~ 600 of these [10]. It is important that, in both [10] and in the work of Girardi *et al.* [5], the cluster ages were derived using the isochrones of Bertelli *et al.* [11], based on stellar-evolution models that take into account mass loss by the most massive stars and moderate convective overshooting [12].

Thus, extensive and homogeneous new material has appeared, enabling the revision of the period–age relation for Cepheids in clusters. This is surely needed, since the old study by Becker *et al.* [13] remains the only one in which a theoretical period–age relation was obtained by comparing models for stellar pulsation and stellar evolution.

3. THE OBSERVATIONAL PERIOD–AGE RELATION

We aimed to impose the strictest possible selection of the cluster Cepheids, based initially on only one criterion—the distance from the cluster center. A direct comparison of the cluster ages and the periods

Table 1. Cepheids in NGC 1943 = SL 430 = LMC 0411

OGLE SC21	*D*	*P*	EROS	GCVS-V	OGLE SC5
1	2	3	4	5	6
187792	0.0004	4.5298	3114		
	0.0006	3.2322	3056		12
187849	0.0010	3.2801	3064		
187853	0.0011	3.3542			63
187788	0.0016	2.8141	3024		
187786	0.0020	3.8599	3085		
187856	0.0020	3.1673	3052		16
187797	0.0048	2.9730	3036		19
187840	0.0051	2.0150	4055*		
187818	0.0181	3.0704		V2287	
187842	0.0213	2.8791			
	0.0403	2.7510			93
	0.0415	2.6685			94

* FO type.
The columns give each Cepheid's (1) designation in OGLE field SC21, (2) distance from the cluster center in degrees, (3) period in days, (4) designation in the EROS project, (5) designation in GCVS-V, and (6) designation in OGLE field SC5 (partially overlapping field 21).

of Cepheids found by OGLE [8] and located within the areas covered by the clusters shows only a very poor correlation. This is naturally explained by the high density of the LMC bar, which means that many field stars are found within the cluster area, and by the fact that the faint clusters discovered during the OGLE project were very poor and had uncertain age determinations. Of the Cepheids suspected as cluster members in [8], we decided to use only those belonging to clusters containing at least two Cepheids, with one of them situated within 0.5 of the cluster radius from the cluster center.

Tables 1 and 2 present data on the Cepheids in the two rich clusters NGC 1943 and NGC 1958, located in the densest part of the LMC bar. Some Cepheids in these clusters were discovered earlier by the EROS project [9], and we have identified these stars with those found in the OGLE project [8] and contained in the General Catalog of Variable Stars [14] (GCVS-V). Tables 1 and 2 show that the OGLE data are the most complete. The clusters display a large difference in their concentrations of Cepheids; we will return to this problem below.

Further, using the AstroView code developed by E.Yu. Efremov (the same code used to reduce the OGLE data), we again compared the positions of Cepheids included in the GCVS-V [14] and the positions of clusters with integrated magnitudes and color indices in [4]. Nearly all the Cepheids identified in this way with positions in or near clusters, were previously known [2]; this is no surprise, since no recent searches for variable stars were carried out in the LMC, apart from inside the bar and a limited number of clusters.

In the case of clusters located outside the LMC bar, and therefore not included in the OGLE project, we limited our analysis to Cepheids located within 0.02° ($\sim$ 18 pc) of the cluster centers. We adopted the data for these Cepheids from [7] and [15], and, for those not present in these papers, from the GCVS-V [14]. We did not include Cepheids with small pulsation amplitudes and nearly sinusoidal light curves (subtype Cs), since their evolutionary status is not entirely clear and they obviously pulsate in the first overtone (see below). Such stars are designated FO (first overtone) in the OGLE catalog. The data on the

Table 2. Cepheids in NGC 1958 = SL462 = LMC 0457

OGLE SC4	D	P	EROS	GCVS-V
1	2	3	4	5
53518	0.001	5.5668	–	
53527	0.002	7.6846	–	
53528	0.002	9.3770	–	
53796	0.005	1.2950*	3002	
53463	0.005	5.3955	3132	
53546	0.012	6.4060	3150	V2532
53514	0.013	5.9755	3145	V2524
176266	0.014	8.8019	3170	
176263	0.019	6.8138	–	V2558
53468	0.020	5.9356	3144	V2522
176307	0.021	2.9763**	4085	
53552	0.023	5.2258	3130	
53458	0.026	6.6196	3155	V2519
44867	0.035	7.1818	–	V2542
45203	0.041	1.1591	–	

* DM type (a double-mode Cepheid).
** FO type.
The columns give each Cepheid's (1) designation in OGLE field SC4, (2) distance from the cluster center in degrees, (3) period in days, (4) designation in the EROS project, and (5) designation in GCVS-V.

stars and clusters we used to derive the period–age relation are collected in Table 3.

Two clusters located in the LMC bar and containing Cepheids deserve special mention. According to [16], two Cepheids are present within 0.02° of the center of the double cluster NGC 1850, which pulsate with the fundamental periods $P = 18.66$ days (No. 17) and $P = 8.56$ days (No. 58). The positions occupied by these stars in the cluster diagram obtained in [16] show that the first Cepheid may belong to a small young cluster located near the edge of the core of NGC 1850 (this young cluster is designated NGC 1850A in [4]), whereas the second Cepheid belongs to the older cluster NGC 1850 itself; we used the ages of the corresponding clusters for each of these stars.

The age of NGC 1958 determined from the cluster's integrated color indices (more specifically, from the parameter S derived from the color indices using Fig. 10 in [5]) is 400 million years, which strongly disagrees with the periods of the Cepheids concentrated toward the cluster and with the age determined from isochrones (80–125 million years). Examining Fig. 10 in [5], we find that the cluster is located in a region of the two-color diagram in which the value of S is very ambiguous, since the value of $B-V$ is anomalously large for the corresponding $U-B$. The color–magnitude diagram for NGC 1958 [17] shows that this anomaly may be due to the presence of an unusually large number of red giants, some of which are probably field stars; the presence of these stars also leads to problems in the derivation of the cluster age from this diagram. However, the Cepheids near NGC 1958 are concentrated towards the cluster center, and the three stars in the cluster core have periods in the same range as those in the cluster corona. In general, the distributions of Cepheids in the general vicinity of the cluster and of their periods do not contradict the presence of Cepheids in the cluster itself as well (Fig. 2). According to the data of [17], the age of NGC 1958 is from 80 to 100–125 million years,

Table 3. Cluster ages and Cepheid periods used to derive relations (1)–(4)

Clusters present in OGLE catalogs [8, 10]				Clusters not included in the OGLE project		
OGLE	Identification	$\log P$ [days]	$\log T$ [years]	Cluster name	$\log P$ [days]	$\log T$ [years]
LMC0512	SL504	0.732	8.1	NGC1755	0.80	8.0
LMC0715	NGC2111	0.454	8.2	NGC1756	0.34	8.4
LMC0715		0.472	8.2		0.43	8.4
LMC0321	HS213	0.618	8.1		0.54	8.4
LMC0321		0.517	8.1	NGC1767	1.48	7.2
LMC0394	NGC1938	0.490	8.4	NGC1850	0.93	7.9
LMC0394		0.529	8.4	NGC1850A[3]	1.27	7.46
LMC0411	NGC1943	0.509	8.08	NGC1866	0.542	8.25
		0.586	8.08		0.435	8.25
		0.449	8.08		0.492	8.25
		0.656	8.08		0.470	8.25
		0.472	8.08		0.497	8.25
		0.516	8.08		0.547	8.25
		0.525	8.08		0.422	8.25
		0.500	8.08		0.506	8.25
LMC0457	NGC1958[1]	0.705		NGC2010	0.455	8.2
		0.745			0.540	8.2
		0.885		NGC2031	0.487	8.2
		0.967			0.646	8.2
		0.807			0.598	8.2
		0.944			0.535	8.2
LMC0585	SL574	0.632	8.05		0.521	8.2
LMC0622	NGC2046	0.505	8.0		0.481	8.2
LMC0622		0.560	8.0		0.496	8.2
LMC0633	NGC2058	0.728	7.94		0.515	8.2
		0.561	7.94		0.470	8.2
		0.523	7.94		0.521	8.2
		0.618	7.94		0.450	8.2
		0.697	7.94		0.505	8.2
		0.669	7.94		0.473	8.2
		0.725	7.94	NGC2136	0.883	8.05
LMC0636	NGC2059[2]	0.746	8.1		1.020	8.05
LMC0648	NGC2065	0.683	8.1	NGC2157	0.889	8.05
		0.642	8.1	NGC2214	1.035	7.9
		0.606	8.1	SL106	1.34	7.7
		0.762	8.1	SL234	1.11	7.9
		0.526	8.1			
		0.561	8.1			

[1] The age is 400 million years according to [4, 5] and 100 million years according to [10].
[2] NGC 2059 was erroneously identified with LMC0632 in [10].
[3] The average of the ages presented in [4, 5] and in [16] is given.

with the latter value being obtained with the same isochrones of Bertelli *et al.* [11] that were used by Girardi *et al.* [5] to calibrate the integrated cluster colors based on their ages. This value agrees well with the age obtained in the OGLE analysis [10]. Thus, there is a firm basis to reject the cluster age derived from S or to exclude this cluster from consideration, based on it's location in the very densest region of the LMC bar. We can now present the least-squares fits of the empirical period–age relations obtained for various determinations of the cluster ages.

(1) We obtain using the ages derived from the integrated colors, adopting for the ages of clusters in the bar the average of the integrated-color and the OGLE ages (except for NGC 1958, whose age was taken directly from the OGLE data [10])

$$\log T = 8.531 - 0.683 \log P. \qquad (1)$$
$$\pm 37 \qquad \pm 57$$

(2) Performing the same fit excluding NGC 1958 yields

$$\log T = 8.546 - 0.715 \log P. \qquad (2)$$
$$\pm 38 \qquad \pm 61$$

Solutions (1) and (2) are virtually identical, since the logarithmic periods for the Cepheids in NGC 1958 are close to the middle of the range for our sample of LMC Cepheids.

(3) The relation changes somewhat if we exclude both NGC 1958 and NGC 1850A (and the 18-day Cepheid that may be associated with this latter cluster), and also restrict our analysis to rich clusters, using only the ages derived from the integrated color indices:

$$\log T = 8.493 - 0.633 \log P. \qquad (3)$$
$$\pm 45 \qquad \pm 72$$

This most "cautious" solution excludes not only NGC 1850A and NGC 1958, but also several poor clusters in the bar that were discovered in the OGLE project. However, it makes use of the OGLE data on Cepheids in rich clusters of the bar.

(4) Finally, another cautious solution uses the same data as solution (1), but does not include NGC 1850A:

$$\log T = 8.509 - 0.643 \log P. \qquad (4)$$
$$\pm 36 \qquad \pm 59$$

These four solutions are displayed in Fig. 3, which shows that they depend rather strongly on the small number of stars with the longest periods. However, Fig. 4 demonstrates that, in contrast to NGC 1850A, the cluster membership of two other long-period Cepheids is beyond doubt. The Cepheid V0333, which has a period of 30.3 days, is located between

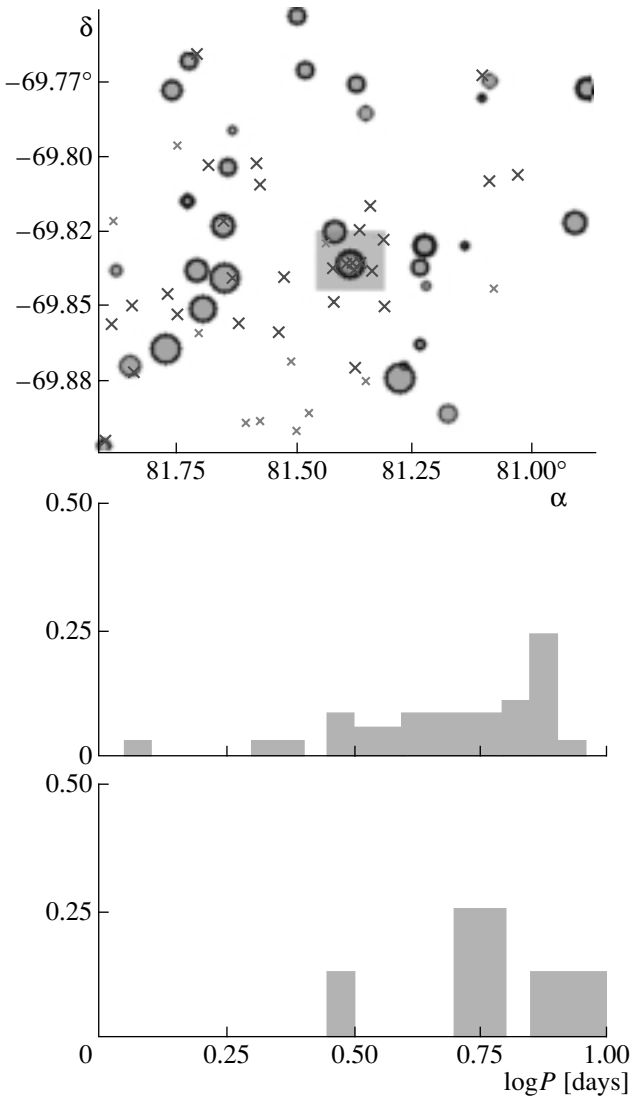

Fig. 2. Top: map of part of the LMC, with the positions of Cepheids (crosses) and clusters (circles) in the vicinity of NGC 1958 indicated. Middle: period distribution of Cepheids in this region, excluding the immediate vicinity of NGC 1658 (shaded grey on the map). Bottom: period distribution for the Cepheids close to NGC 1958 (in the grey-shaded area). The distributions are normalized, and the logarithms of the period are plotted. The Cepheids concentrated towards the cluster and located in its general vicinity have similar period (age) distributions, but the period range in the cluster is narrower. The left "tails" of the distributions correspond to FO Cepheids. (This figure was plotted using the AstroView code developed by E.Yu. Efremov.)

the clusters NGC 1767 (=SL 123) and SL 123, which have almost equal ages; V1040 is within 0.3° of these clusters, at the edge of SL 106. This same region, corresponding to the OB association LH8 [2], contains two other rich young clusters and the Cepheid V0432, with a period of 36.8 days. Note

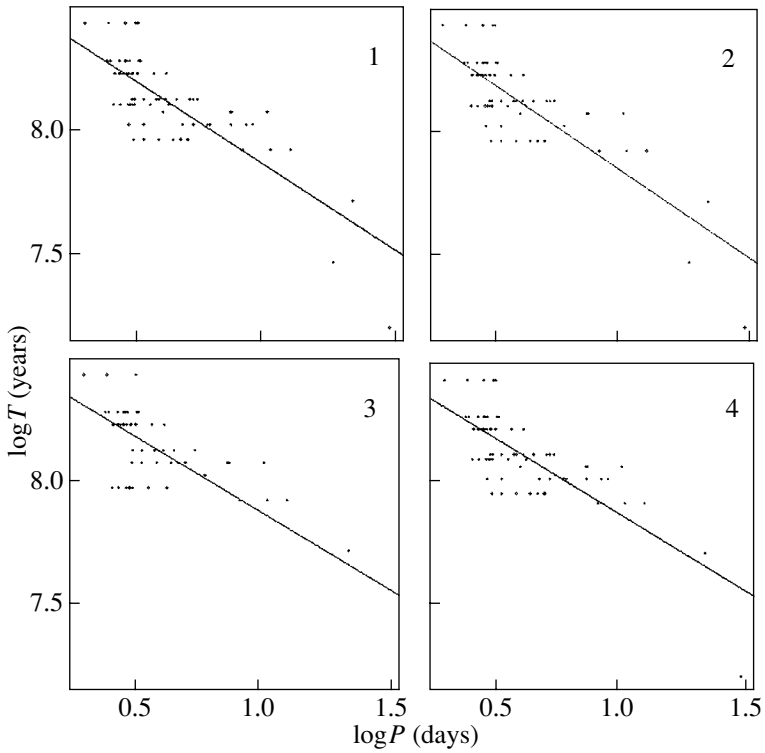

Fig. 3. Different versions of the period–age relations (marked with numbers on the diagrams, see text for details). The logarithms of the Cepheid periods and cluster ages are plotted.

that the 33-day Cepheid H40, which is located in the outskirts of the rich cluster vdB0 in M 31, whose age from integrated UBV photometry [18] is about 35 million years, shows a good agreement with all four relations.

We can now compare our results with those recently obtained by other authors who have used the same age scale for the clusters. Using the most reliable LMC stars and cluster ages derived from integrated photometry, Efremov and Elmegreen [15] obtained the period–age relation

$$\log T = 8.492 - 0.509 \log P, \qquad (5)$$

close to our relation (3).

Based on ages derived from the Geneva isochrones using the color–magnitude diagrams of nine rich LMC clusters and seven open clusters in our Galaxy containing 37 Cepheids, Grebel and Brandner [19] obtained the relation

$$\log T = 8.5454 - 0.7302 \log P, \qquad (6)$$
$$\pm 8 \qquad \pm 5$$

close to our relation (2). Unfortunately, the details of their study were not published. The low uncertainties in this determination may provide evidence that differences in the chemical compositions of young stars in the Galaxy and in the LMC (whose metal abundances are a factor of two to three lower than those of the Milky Way) influence the parameters of the period–age relation only weakly. In any case, we can conclude that the "empirical" period–age relation based on a modern age scale for the clusters that takes into account moderate convective overshooting is established with reasonable certainty.

4. THEORETICAL DETERMINATIONS OF THE PERIOD–AGE RELATION

No theoretical determinations of the period–age relation based on modern evolutionary models for massive stars have been published, but we can obtain a rough estimate from the models of Bono *et al.* [20] for the chemical composition $Y = 0.255$ and $Z = 0.01$, which approximately corresponds to that of the LMC. For this purpose, we adopted the sums of the lifetimes of massive stars in the stages of hydrogen and helium burning from Table 2 of [20] and compared these with the periods of Cepheids (of the corresponding masses) presented in Table 7 of that same paper for the stage of the third crossing of the instability strip, in which most Cepheids are located. This last crossing (according to [20]) should correspond to an age close to the end of helium burning. This yielded the theoretical period–age relation

$$\log T = 8.471 - 0.748 \log P. \qquad (7)$$

This relation is close to the above empirical relations, testifying at the very least to the self-consistency of the computations of the evolution and pulsation of massive stars presented in [20]. Note that the analysis of [20] used the classical criterion for the boundary of the convective zone, whereas we have used cluster ages obtained taking into account moderate convective overshooting. Apparently, the approach used in [20] to compute the evolution of massive stars compensated for the increase in the ages of massive stars in models with convective overshooting. As is noted in [20], the question of whether it is necessary to take this overshooting into account remains open. The observational data, even for such a rich cluster as NGC 1866, likewise do not permit us to draw definite conclusions [21].

The period–age relation we derived a quarter of a century ago [2] (based mainly on LMC stars) was

$$\log T = 8.16 - 0.68 \log P, \qquad (8)$$

while the theoretical relation obtained at that same time by Becker *et al.* [13] for $X = 0.70$ and $Y = 0.02$ can be written [13]

$$\log T = 8.45 - 0.78 \log P. \qquad (9)$$

Of course, we cannot completely rule out the possibility that the excellent agreement between relations (9) and (7) and the good agreement between (9) and (1)–(4) have come about by chance; however, as a minimum, we must reject both (8) and the relation

$$\log T = 8.70 - 0.53 \log P, \qquad (10)$$

derived in [22] after reestimating the ages for LMC clusters using the first models with overshooting, which led to a three-fold decrease in the ages of young clusters.

The best hope for verifying the age scale of massive stars derived from stellar-evolution theory is to compare the corresponding ages to age estimates derived from stellar-dynamical data in some way. We noted earlier [23, p. 156] that the period–age relation (10) was incompatible with the gradient of Cepheid periods (ages) detected in the S4 arm of the Andromeda galaxy, since it implied that this part of the arm was close to corotation. Recently, Dambis [22] also found that the short age scale for young clusters that follows from models with full convective overshooting was in poorer agreement with the vertical motions of stellar clusters and with density estimates for the disk of our Galaxy.

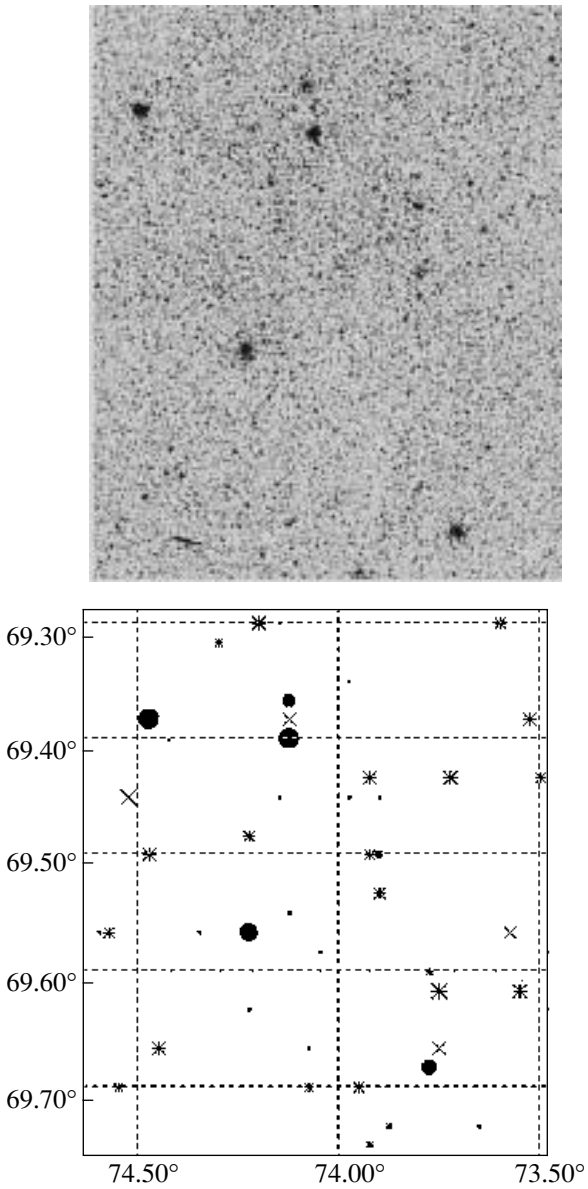

Fig. 4. Region of the association LH8 containing long-period Cepheids (crosses) and young clusters (circles). The image is from the DSS, the map (on the same scale) was plotted using data from the GCVS-V [14] and [4] and the AstroView code. NGC 1767 and SL 123 are towards the top, near the middle of the image.

5. REMARKS ON CEPHEID EVOLUTION

Clusters containing several Cepheids, including some Cs Cepheids, enable us to investigate the evolutionary stage of this subtype of Cepheid. Cs Cepheids have almost sinusoidal light curves and small pulsation amplitudes. They also have lower temperatures than ordinary Cepheids. Many years ago, we suggested that Cs stars are in the stage of the rapid first crossing of the instability strip after leaving the main sequence [24].

Some stars of this type indeed show rapid period increases. Their period distribution is biased towards shorter periods, consistent with the fact that the loops of the evolutinary tracks on which most Cepheids should be situated begin to penetrate (and, for longer

Table 4. Cepheids in NGC 2058 = SL 614 = LMC 0633

OGLE SC16	D	P	Mode	V	I	GCVS-V
1	2	3	4	5	6	7
172383	0.001	5.3392	FU	15.59	14.78	V3628
172455	0.001	1.9726	FO	15.89	15.29	
172460	0.001	3.3406	FU	15.99	15.25	
172450	0.002	2.1231	FO	15.96	15.31	
172459	0.003	2.0879	FO	15.89	15.28	
177774	0.003	4.9764	FU	15.50	14.76	
177777	0.003	4.6769	FU	15.48	14.79	
177823	0.004	1.9179	FO	16.03	15.42	
172447	0.005	3.6438	FU	15.83	15.15	
172435	0.007	2.2710	FO	15.96	15.28	
177773	0.010	4.1465	FU	15.93	15.05	
177781	0.012	5.3454	FU	15.58	14.80	
235480	0.012	2.0455	FO	16.04	15.37	

The columns give each Cepheid's (1) designation in OGLE field SC16, (2) distance from the cluster center in degrees, (3) period in days, (4) pulsation mode, (5) mean V magnitude, (6) mean I magnitude, (7) designation in the GCVS-V.

periods, cross) the instability strip only for masses corresponding to periods of about 3 days or longer.

However, the discovery of numerous Cs Cepheids in the LMC bar during the OGLE project [25] represented strong evidence against this hypothesis. The theory predicts that the first instability-strip crossing should be an order of magnitude shorter than subsequent crossings, in agreement with the fraction of Cs stars among Cepheids in the Galaxy; however, there are only a factor of two fewer Cs stars than ordinary Cepheids in the LMC. Like most other authors, Udalski *et al.* [25] consider low-amplitude stars to be pulsating in the first overtone, and designate them FO. In the $\log P$ range from 0.5 to 0.6, there are four times as many FO Cepheids as stars pulsating with the fundamental period (FU) (cf. Fig. 6 in [25]). Udalski *et al.* [25] also confirmed that the luminosity of FO stars with a given period is higher than that of FU stars by a factor that exactly corresponds to the ratio of the first-overtone period to the fundamental period (0.72). Thus, the problem of the nature of Cs Cepheids could be considered essentially solved; however, there remain some puzzling data.

In LMC clusters, the Cepheids classified as FO stars in the OGLE project also have the shortest periods in a given cluster. If their masses were all approximately equal, the difference of the periods of the stars pulsating in the first overtone should have disappeared after multiplying these periods by 1/0.72—the ratio of the fundamental period to the first-overtone period. But this is not the case. The periods of the majority of the FO stars remain the shortest in their clusters after this "reduction to the FU period." It is tempting to explain this by suggesting that the loops of the evolutionary tracks of lower-mass cluster stars do not yet penetrate the instability strip.

Thus, in the first-crossing hypothesis, the difference in the periods of the ordinary (FU) and Cs (FO) Cepheids could be explained if the latter have systematically lower masses, so that Cs stars are at an earlier stage of evolution. Then, the difference in the ages of the FO and FU stars in a single cluster should correspond to the difference between the times their progenitors left the main sequence minus the time spent by the higher-mass Cepheids in the red supergiant phase, after the first and second crossings. In this way, we can compare the evolution scales on the main sequence and after leaving it [26]. Of course, this assumes that all the stars in a cluster form nearly simultaneously, at least within some range of similar masses. The observed dispersion of the periods of the FU Cepheids in a single cluster usually does not exceed 0.2 dex, in agreement with this hypothesis, since this value is approximately equal to the theoretical difference between the Cepheid periods for different instability-strip crossings.

Note also that, in NGC 2058 (Fig. 5), which contains a very large number of FO/Cs stars (six), the periods of these stars increase, on average, with increasing color index, as is expected if they are moving from left to right in the color–luminosity diagram at a rate determined by their mass (period). We can see from Table 4, which contains data on the Cepheids in NGC 2058, that the positions of the FU Cepheids in the color–magnitude diagram trace a continuation of their evolutionary track, which is obviously possible only if the age dispersion for the massive stars in the cluster is small. Moreover, this part of the track generally resembles the corresponding part of the theoretical evolutionary track for a $5M_{\odot}$ star with $Y = 0.255$ and $Z = 0.01$ presented by Bono *et al.* [20], whereas it follows from Table 7 of [20] that the Cepheid periods for these parameters are 1.8 days for the first and 3.4 days for the second crossing, close to the observed periods of the FO and FU Cepheids in NGC 2058. However, there are equal numbers of these two types of Cepheids in the cluster, whereas the theoretical durations of the corresponding instability-strip crossings differ by an order of magnitude. If the theory is unable to considerably increase the duration of the first instability-strip

crossing, we can propose that all Cepheids pulsate in the first overtone during the first crossing, although most FO stars are in different stages of evolution. This also agrees with the fact that far from all Cs stars demonstrate rapid period increase, but nearly all such stars have low amplitudes and sinusoidal light curves.

6. CONCENTRATION OF CEPHEIDS IN CLUSTERS

Finally, let us briefly discuss differences in the concentrations of Cepheids towards the centers of their clusters for clusters for which exhaustive studies have been carried out. What could be the origin of such differences for clusters with approximately the same ages and masses? They are especially strong for NGC 1958 (Fig. 6) and NGC 1943 (Fig. 7), in the center of the LMC bar. Recall that the stars in question are massive (3–12 $M_\odot$), and that Cepheids (along with red supergiants) are the most evolved and hence the most massive stars in their parent clusters. Dynamical segregation by mass should lead to their concentration in the cluster cores. This is often, but not always, the case. Too many Cepheids are encountered in the coronas of clusters in our Galaxy [1], and this is also true for the LMC.

Turner [27] noted that the number of Cepheids in the coronas of Galactic clusters was a factor of 1.4–2.4 higher than in the cluster cores, and explained this by suggesting that, in the cluster cores, there could be many binaries that were so close that their components could not reach the size of supergiant Cepheids. Another explanation is also possible, however. Brandl and Chernov [28] note that, in addition to being concentrated toward the center, the most massive stars of R136, the central cluster of 30 Dor, are also present 2–3 pc from the center. They suggested that the most probable explanation is the ejection of such corona stars from the dense core during dynamical interactions between the most massive stars, which were initially rapidly concentrated in the cluster core.

This same mechanism may be responsible for the high fraction of Cepheids, and probably also of red supergiants, in cluster coronas, as is observed, for example, for NGC 2100 in the LMC. The degree of concentration of massive stars towards cluster centers varies strongly in different LMC clusters [29, 30]. For example, in their study of the structure of six rich LMC clusters, de Grijs *et al*. [30] found that they could be subdivided into three pairs, with the ages, metallicities, and distances from the center of the LMC being the same for each pair, but with the core radii of the two clusters in a pair being very different. We observe essentially the same effect in the pair consisting of NGC 1958 and NGC 1943. de Grijs *et al*. [30] concluded that the differences between the initial mass functions of the two clusters in the pairs they studied were not large, and could not give rise to the large differences in their core radii. It is possible that these pairs, like our clusters with Cepheids, were caught at different stages of their dynamical evolution that are characterized by different core densities. Depending on the initial conditions (first of all, the initial fraction of binary systems [31, 32]), these stages could well be different even for clusters with similar masses and ages. It was shown in [31] that the dynamical evolution of a cluster, accompanied by the departure of up to half of the massive stars from the cluster core, could last from 10 to 50 million years for clusters containing 1000–10 000 stars. It seems

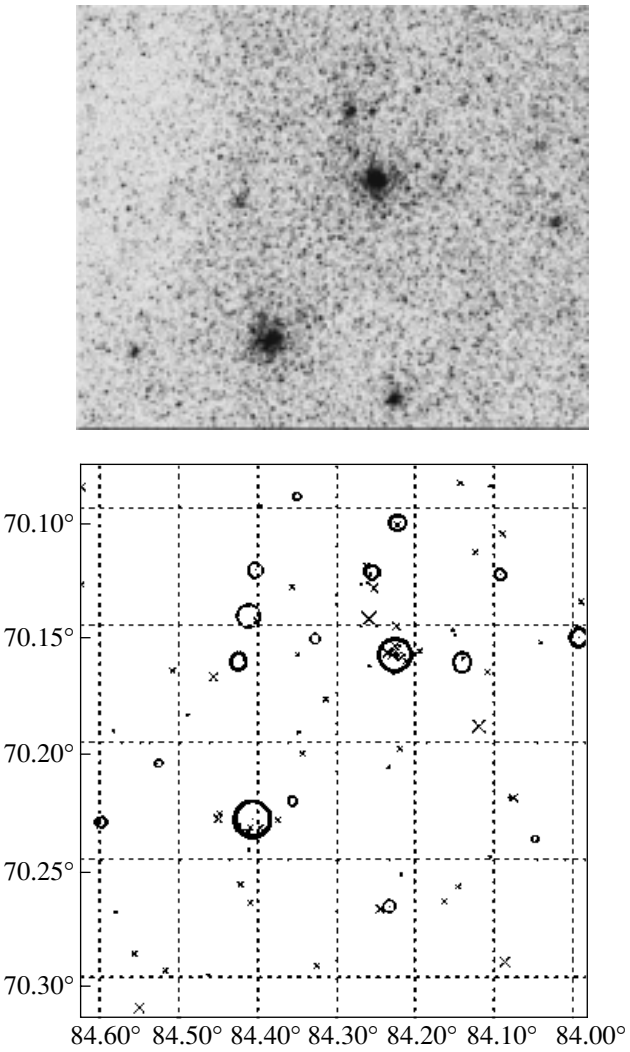

Fig. 5. Cepheids (crosses) and clusters (circles) near the eastern end of the LMC bar. NGC 2058 is to the upper right and NGC 2065 to the lower left. Cs Cepheids are plotted as small crosses. The image is from the DSS, and the map was plotted using the OGLE data with the AstroView code.

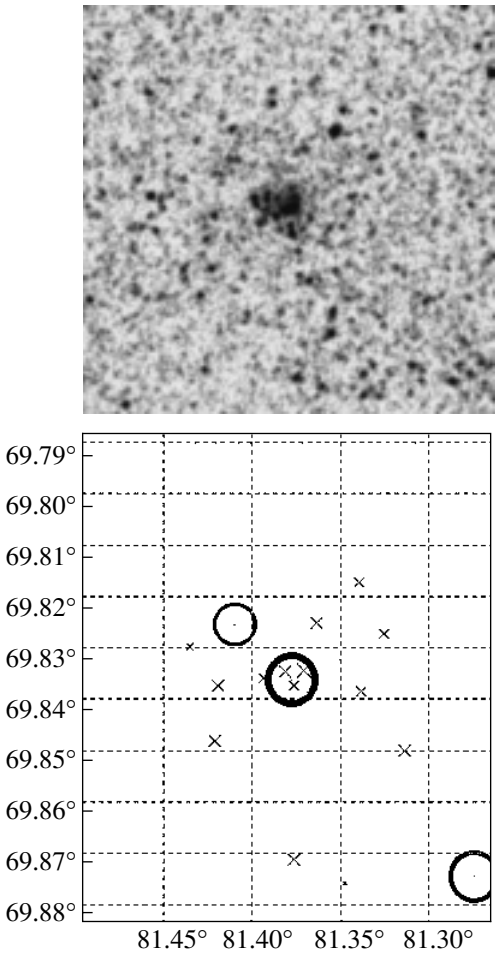

Fig. 6. Positions of Cepheids (crosses) in NGC 1958 (in the center) and in its nearest vicinity. The image is from the DSS, and the map was plotted using OGLE data with the AstroView code.

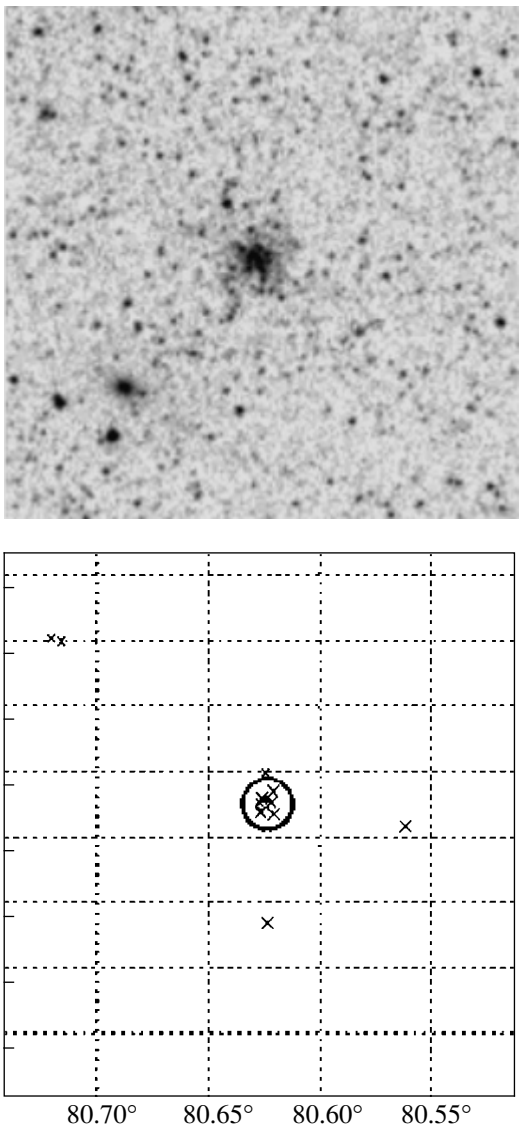

Fig. 7. Concentration of Cepheids towards NGC 1943. The image is from the DSS, the map was plotted using OGLE data with the AstroView code.

very likely that precisely such differences in the rates of the dynamical evolution can explain the strong central concentration of Cepheids toward the core of NGC 1943, slightly less strong concentration in NGC 2058, still weaker concentration in NGC 2065, and very weak concentration in NGC 1958 (Figs. 5 and 6). All these clusters are located within the fields studied in the OGLE project, and the completeness of discoveries of their Cepheids can be considered exhaustive (according to [25], it is at least 96%).

Before the OGLE results were published, it was believed that the concentration of Cepheids in the region of NGC 1958 was the result of chance projection; however, three OGLE Cepheids fall exactly in the core region. Nevertheless, apart from the discrepant age determinations for this cluster (see above), the high number of Cepheids in this not very rich cluster remains enigmatic. We suggested that the presence of three Cepheids in the Milky Way's poor cluster NGC 7790 could be understood if its age corresponds to the time when the loop ends of the Cepheids' evolutionary tracks (where the star's motion along the track is the slowest) is within the instability strip [23]. However, the Cepheid periods in NGC 1958 (6–9 days) are much longer than in NGC 7790 (4.5–5 days). This may be due to differences in chemical composition that influence the positions of the track loops; alternatively, it may be that, in the case of the Cepheids in NGC 1958, which is located in the densest part of the LMC bar, we really are dealing with a chance projection. The structures and chemical compositions of the LMC clusters containing several Cepheids each, as well of those having no Cepheids at all, despite their suitable age and richness, deserve a detailed study.

7. CONCLUSIONS

The period–age relation for Cepheids based on their membership in stellar clusters can be considered quite firmly established. The theoretical relation needs further improvements. It is possible that both the chemical composition and differences in the details of the evolutionary-track computations for massive stars influence this relation only weakly. The reason for this is not clear; some sort of mutual compensation of times and periods is possible. Further refinements of the evolutionary tracks of massive stars and their lifetimes within the instability strip are also needed; the period distribution of Cepheids cannot be correctly translated into their age distribution without normalization to these lifetimes.

The numerous Cepheids in rich LMC clusters provide the best data for comparing theories of stellar evolution and pulsation with observational results. It would be especially interesting to distinguish which of the observed characteristics of Cepheids (apart from evolutionary variations of their periods [33]) might characterize particular crossings of the instability strip. The large relative number of Cs Cepheids contradicts the hypothesis that these stars are undergoing their first crossing, though the characteristics of the Cepheids in NGC 2058, with half of its Cepheids belonging to this type, are consistent with this hypothesis. Nevertheless, it is difficult to imagine why the theory would predict a rate of stellar evolution during the first crossing that was much too high. It is possible that all Cepheids undergoing their first crossing are first-overtone pulsators, whereas the converse is not true: most FO stars are actually in other evolutionary stages. This is also consistent with the fact that not all Cs stars demonstrate a rapid evolutionary increase in their periods [33]. Long series of observations of the sort needed to establish the presence of evolutionary period variations have been accumulated for only a few stars in our Galaxy. Frequently, claims of period variations based on differences between derived periods are unfounded, as the differences are due purely to uncertainties in periods derived from a short series of observations.

It is important to understand why Cepheids are not present in several rich clusters of suitable ages. New, even more careful searches for Cepheids in such clusters may be needed. The differences in the concentration of Cepheids toward cluster centers may come about because the clusters are in different stages of their dynamical evolution. This hypothesis agrees with recent theoretical results indicating that massive stars can initially rapidly settle toward the cluster center, after which some of them may be ejected into the cluster corona in the course of dynamic interactions between stars in the dense cluster core. The core density and the efficiency of stellar dynamic interactions can vary rapidly and repeatedly due to gravithermal oscillations that arise during the cluster's evolution [34].

It is an interesting possibility that Cepheids, whose ages and masses can be determined with the highest accuracy achievable for single stars, can be used not only to reconstruct the star-formation history of stellar complexes, but also as test particles in studies of the dynamic evolution of clusters.

ACKNOWLEDGMENTS

The author is grateful to the referee, N.N. Samus', for many useful remarks concerning the first version of this paper and to E.Yu. Efremov for his assistance during the preparation of the figures. This study used data from the NASA Astrophysical Data System and the Digital Sky Survey (DSS). It should be specially noted that we based our research on the results of the OGLE project, which have been published with an exhaustive completeness that is far from common. This study was supported by the Russian Foundation for Basic Research (project no. 03-02-16288) and a grant from the Program of Support to Leading Scientific Schools of Russia (NSh 389.2003.2).

REFERENCES

1. Yu. N. Efremov, Perem. Zvezdy **15**, 242 (1964).
2. Yu. N. Efremov, Astron. Zh. **55**, 272 (1978) [Sov. Astron. **22**, 161 (1978)].
3. Yu. N. Efremov, Astron. Zh. **79**, 879 (2002) [Astron. Rep. **46**, 791 (2002)].
4. E. Bica, J. J. Claria, H. Dottori, *et al.*, Astrophys. J., Suppl. Ser. **102**, 57 (1996).
5. L. Girardi, C. Chiosi, G. Bertelli, and A. Bressan, Astron. Astrophys. **298**, 87 (1995).
6. M. Mateo, E. W. Olszewski, and B. Madore, Astrophys. J. **353**, L11 (1990).
7. M. Mateo, Publ. Astron. Soc. Pac. **104**, 824 (1992).
8. G. Pietrzynski and A. Udalski, Acta Astron. **49**, 437 (1999).
9. C. Afonso, J. N. Albert, C. Alard, *et al.*, astro-ph/9907355 (1999).
10. G. Pietrzynski and A. Udalski, Acta Astron. **50**, 337 (2000).
11. G. Bertelli, A. Bressan, and C. Chiosi, Astron. Astrophys., Suppl. Ser. **105**, 275 (1994).
12. F. Fagotto, A. Bressan, G. Bertelli, and C. Chiosi, Astron. Astrophys., Suppl. Ser. **105**, 29 (1994).
13. S. A. Becker, I. Iben, and R. S. Tuggle, Astrophys. J. **218**, 633 (1977).
14. *General Catalogue of Variable Stars,* Ed. by N. N. Samus' (Kosmoinform, Moscow, 1995) [in Russian], Vol. 5.
15. Yu. N. Efremov and B. G. Elmegreen, Mon. Not. R. Astron. Soc. **299**, 588 (1998).
16. K. M. Sebo and P. R. Wood, Astrophys. J. **449**, 164 (1995).

17. V. Ripepi, F. Bauer, and J. P. Beaulieu, in *IAU Symp. 190: New Views of the Magellanic Clouds,* Ed. by Y. H. Chu *et al.* (1999), p. 559.
18. S. van den Bergh, Astrophys. J., Suppl. Ser. **171**, 145 (1969).
19. E. K. Grebel and W. Brandner, *The Magellanic Clouds and Other Dwarf Galaxies*, Ed. by T. Richtler and J. M. Braun (Shaker, Bonn, 1998), p. 151.
20. G. Bono, F. Caputo, S. Cassisi, *et al.*, Astrophys. J. **543**, 955 (2000).
21. E. Brocato, V. Castellani, E. Di Carlo, *et al.*, astro-ph/0302458 (2003).
22. A. K. Dambis, astro-ph/0303461 (2003).
23. Yu. N. Efremov, *Star Formation Regions in Galaxies* (Fizmatlit, Moscow, 1989) [in Russian], p. 53.
24. Yu. N. Efremov, Perem. Zvezdy **16**, 365 (1968).
25. A. Udalski, I. Soszynski, M. Szymanski, *et al.*, Acta Astron. **49**, 223 (1999).
26. Yu. N. Efremov and I. M. Kopylov, Izv. Krymsk. Astrofiz. Obs. **36**, 240 (1967).
27. D. G. Turner, *IAU Coll. 82: Cepheids: Theory and Observations* (1985), p. 209.
28. B. R. Brandl, D. F. Chernoff, and A. F. J. Moffat, *IAU Symp. 207: Extragalactic Star Clusters*, Ed. by D. Geisler, E. K. Grebel, and D. Minniti (ASP, San Francisco, 2002), p. 694.
29. R. de Grijs, G. F. Gilmore, R. A. Johnson, and A. D. Mackey, Mon. Not. R. Astron. Soc. **331**, 245 (2002).
30. R. de Grijs, G. F. Gilmore, A. D. Mackey, *et al.*, Mon. Not. R. Astron. Soc. **337**, 597 (2002).
31. P. Kroupa, astro-ph/0001202 (2000).
32. R. Spurzem, Astron. Astrophys. Trans. **20**, 55 (2001).
33. L. N. Berdnikov, V. V. Ignatova, J. A. R. Caldwell, and C. Koen, New Astron. **4**, 625 (2000).
34. M. Giersz and R. Spurzem, astro-ph/0301643 (2003).

Translated by N. Samus'

Astronomy Reports, Vol. 47, No. 12, 2003, pp. 1013–1026. Translated from Astronomicheskiĭ Zhurnal, Vol. 80, No. 12, 2003, pp. 1099–1113.
Original Russian Text Copyright © 2003 by Nazarenko, Glazunova.

Hydrodynamical Modeling of Mass Transfer in the Close Binary System β Lyr

V. V. Nazarenko and L. V. Glazunova

Astronomical Observatory of Odessa, National University, Odessa, Ukraine

Received December 15, 2002; in final form, March 14, 2003

Abstract—We present results of two-dimensional hydrodynamical simulations of mass transfer in the close binary system β Lyr for various radii of the accreting star and coefficients describing the interaction of the gaseous flow and the main component (primary). We take the stellar wind of the donor star into account and consider various assumptions about the radiative cooling of the gaseous flow. Our calculations show that the initial radius of the flow corresponding to our adopted mass-transfer rate through the inner Lagrange point (L_1) of $(1-4)\times10^{-5}M_\odot$/yr is large: 0.22–0.29 (in units of the orbital separation). In all the models, the secondary loses mass through both the inner and outer (L_1 and L_2) Lagrange points, which makes the mass transfer in the system nonconservative. Calculations for various values of the primary radius show a strong dependence on the coefficient f_v that models the flow–primary interaction. When the radius of the primary is 0.5, there is a strong interaction between the gas flow from L_1 and the flow reflected from the primary surface. For other values of the primary radius (0.1 and 0.2), the flow does not interact directly with the primary. The flow passes close to the primary and forms an accretion disk whose size is comparable to that of the Roche lobe and a dense circum-binary envelope surrounding both the disk and the binary components. The density in the disk varies from 10^{12} to 10^{14} cm^{-3}, and is 10^{10}–10^{12} cm^{-3} in the circum-binary envelope. The temperature in the accretion disk ranges from 30 000 to 120 000 K, while that in the circum-binary envelope is 4000–18 000 K. When radiative cooling is taken into account explicitly, the calculations reveal the presence of a spiral shock in the accretion disk. The stellar wind blowing from the secondary strongly interacts with the accretion disk, circum-binary envelope, and flow from L_2. When radiative cooling is taken into account explicitly, this wind disrupts the accretion disk.
© 2003 MAIK "Nauka/Interperiodica".

1. INTRODUCTION

β Lyr is a well studied eclipsing binary of spectral type B(6–8)I + B(0–3)V with an orbital period of 12.9 day, discovered by J. Goodricke in 1789. Since that time, thanks to its high brightness, it has been very carefully studied for some 200 years. Observational studies have been especially intense in the last 30 years, in the X-ray, UV, optical, radio, and IR.

β Lyr is usually considered to be a close binary system in which the less massive star (the secondary or donor star) is losing mass through the inner Lagrangian point at a rate close to $10^{-5}M_\odot$/yr, resulting in the formation of an optically thick envelope around the primary. It is possible that this envelope enshrouds the entire system. This envelope is usually considered to have the form of an accretion disk rotating around the primary (the model of Hubeny and Plavec [1]). Currently, there are two models for the envelope of β Lyr. The first was suggested by Wilson [2], and envisions a massive dense disk in the form of a prolate ellipsoid. In the second model, suggested by Hubeny and Plavec [1], the envelope forms an accretion disk whose vertical structure is in hydrostatic equilibrium. In contrast to the model of Wilson, the disk in the Plavec–Hubeny model has negligible mass, obeys Keplerian motion, and has the overall form of a torus. Its height (thickness perpendicular to the orbital plane) grows with distance from the main component. Comparing models with observations and carrying out geometrical modeling of the radiating area of the primary, Hubeny and Plavec [1] concluded that the disk radius is about 25 solar radii and the disk height about 6 solar radii.

Wilson and Terrell [3] critically examined the Hubeny–Plavec model and suggested their own alternative model for the accretion disk. Hubeny *et al.* [4] discussed the criticism of Wilson and Terrell [3] and suggested a strategy for new studies of β Lyr. Harmanec *et al.* [5] proposed that most of the emission in the Hα and HeI λ6678 Å lines forms in jet-like structures moving perpendicular to the orbital plane. Their attempts to model the observed light curve of β Lyr led Linnell *et al.* [6] to conclude that it is not possible to describe the UV and optical observations using a single set of parameters. Recently, Linnell [7] suggested a two-temperature

1063-7729/03/4712-1013$24.00 © 2003 MAIK "Nauka/Interperiodica"

model for the accretion disk that describes the UBV and OAO2 photometric data well.

The generally accepted rate of mass transfer through the inner Lagrangian point in β Lyr is close to $10^{-5} M_\odot$/yr (see, e.g., [7–9]).

Only one paper concerning numerical modeling of the mass transfer in the β Lyr system has been published: Bisikalo *et al.* [9] present the results of three-dimensional hydrodynamical simulations of this mass transfer. Bisikalo *et al.* [9] restricted their computational domain to the Roche-lobe of the primary and took radiative cooling into account implicitly, via variation of the specific heat. These calculations showed that the matter that is transferred from one component to another forms an accretion disk without a hot spot. Moreover, the matter that flows around the primary collides with the original stream, forming a shock wave. A jet-like structure is formed above the orbital plane in the region of the shock, but the velocity of the gas in these jets is much lower than is observed [9]. A torus-like envelope (halo) forms above the accretion disk and primary. These computations also showed that the results were not influenced by the assumed density at L_1.

Thus, some important steps in studies of the processes involved in the mass transfer in β Lyr have been made. However, a more accurate description of the physical processes influencing the mass transfer is necessary if we wish to describe the mass transfer not only qualitatively, but also quantitatively, including the formation of the flow at L_1, radiative cooling, the interaction of the stream with the atmosphere of the primary, the the influence of the stellar wind from the secondary.

The aim of the current paper is to investigate the influence of these physical processes on the formation of gaseous envelopes in the β Lyr system via two-dimensional (2D) hydrodynamical computations, varying the radius of the primary (since it is not known precisely) and the rate of mass transfer.

2. DESCRIPTION OF MASS TRANSFER IN CLOSE BINARY SYSTEMS. INITIAL MODEL. MODEL OF THE BINARY AND BOUNDARY CONDITIONS

Our approach to describing mass transfer in close binary systems is based on two assumptions: (1) instead of applying the model of Lubow and Shu [10], we determine the structure of the stream at L_1 by computing the formation of the stream in the vicinity of L_1; (2) we use the model atmospheres of Kurucz [11] to obtain an initial model for the structure of the secondary in the vicinity of L_1. Our model for the formation of the stream at L_1 is described in detail in [12]. We believe that our approach enables us to accurately take into account the properties of the mass transfer in any close binary. This is especially true of long-period systems like β Lyr, where the donor star has a very large radius, and hence an extended atmosphere, which should, in turn, result in the formation of a stream with a large radius in the vicinity of L_1. The distribution of the pressure gradient in the vicinity of L_1, and hence the radius of the stream at this point, influence the structure of the envelopes that form due to the motion of the stream inside the Roche lobe of the primary and the interaction of the stream with this star. If the stream has a large radius, it may interact with the atmosphere of the primary even if this star does not have a large radius, for example, of the order of 0.1 (throughout, we use units of the semimajor axis of the orbit). The parameters of the stream at L_1 probably likewise influence the dimensions of the stream and envelopes in the z direction.

We shall apply two schemes for taking into account the radiative cooling of the gas—implicit and explicit—and compare the results. The radiative cooling is taken into account implicitly by imitating the radiative losses by reducing the adiabatic index in the equation of state for an ideal gas, as in [9, 13]. We explicitly take into account radiative losses by computing the radiative cooling for an optically thin plasma that is in ionization equilibrium; this method is described in detail by Cox and Daltabuilt [14]. This radiative-cooling model was first used in the hydrodynamical calculations of Blondin *et al.* [15]. We took into account self-absorption in the optically thick gas layer by multiplying the radiative cooling by the optical path in the envelope in the z direction. The interaction of the stream and primary was modeled by computing the fraction of the gas that is accreted by introducing a special coefficient f_v that determines the velocity at the surface of the primary: this velocity is equal to $f_v \times V$, where V is the velocity at the interface with the primary. We used the values $f_v = 0.1$ and 0.5, in order to estimate the influence of this coefficient on the results of the computations.

In our model for the binary system, the components have circular orbits around the common center of mass of the system, a gravitational Roche model is applied, and we assumed synchronous rotation of the components. We also assumed that the gas was ideal and, for simplicity, composed purely of hydrogen.

We used the model atmospheres of Kurucz [11] with $T_{\rm eff} = 10\,000$ K and $\log g = 2.5$ to construct the initial model for the secondary in the vicinity of L_1. We assumed that the contours of equal pressure inside the Roche lobe of the secondary coincide with the equipotential contours of the gravitational field of the binary. Outside the Roche lobe of the secondary, the

contours of equal pressure are parallel to this Roche lobe [12].

We applied the following boundary conditions. At the outer boundary of the computational region, the gradients of physical quantities are equal to zero (enabling a free outflow of gas through this boundary). Deep inside the Roche lobe of the secondary and at its surface, the physical quantities remain equal to their initial values.

The mass of the secondary was set equal to $2M_\odot$, and the mass of the primary to $12M_\odot$ [16]. Two numerical grids were used for the computations. The main grid covers the space in the vicinity of the binary from -1.5 to $+2.5$ (along the x axis) and from -2.5 to $+2.5$ (along the y axis), with the number of cells in this grid being 275×285. The second grid covers a larger space: from -9 to $+17$ along the x axis and from -19 to $+19$ along the y axis, with the number of cells being 475×585. The large grid enables us to follow the motion of gas that departs from the system.

The computations covered 12 to 19 orbital periods, to ensure that a steady state of the system was established (we assumed that the system attains this steady state when the rates of accretion and mass loss by the system become constant in time).

To compute the mass transfer in the close binary, we solved the non-stationary Euler equations using the "large-particle" method of Belotserkovskiĭ and Davydov [17]. We applied a version of the method that uses artificial viscosity in the first step and has second-order accuracy in space and time. The simplicity of the organization of the computations in the large-particle method enables the use of personal computers for the solution of nonstationary problems in the numerical hydrodynamics of close binaries.

3. MODELING OF MASS TRANSFER IN β Lyr AT A RATE OF $10^{-5}M_\odot$/yr

3.1. Computations with Implicit Account of Radiative Cooling

3.1.1. Primary radius equal to 0.5, $f_v = 0.5$ and 0.1. Figure 1 ($f_v = 0.5$) and Figure 2 ($f_v = 0.1$) show contours of equal density and temperature and the velocity field for the steady-state solution obtained for this case. The radius of the stream at L_1 is 0.2 (the model of Lubow and Shu [10] gives a radius of 0.17 for the same mass-transfer rate). Due to the large radius of the primary, the mass transfer results in the Roche lobe of the primary being filled with gas, and the formation of an almost contact binary system. A comparison of Figs. 1 and 2 shows that an almost symmetric envelope forms when $f_v = 0.1$ (90% of the matter is reflected by the primary). When $f_v = 0.5$ (50% of the matter is reflected by the primary and 50% is accreted), an asymmetric envelope forms, with the maximum density occurring at orbital phases 0.65–0.75. In both cases, the stream reflected by the primary interacts with the stream flowing from L_1, resulting in the formation of a high-temperature region between the primary and L_1. The temperature in this area can reach 2×10^5–3×10^5 K.

As we can see in Figs. 1 and 2, none of the shock waves that are characteristic of close binaries with compact objects [18, 19] arise in the computational region. The only exception is shock III (our notation for labeling the shocks coincides with that of Bisikalo *et al.* [18, 19]). The computations also show that the system loses essentially no mass from the stream flowing from L_1. Matter from the secondary flows out through L_2, forming a stream of matter that leaves the system. Thus, the system loses mass mainly through L_2. The rate of mass flow through L_2 is about 40% of the rate of flow through L_1. Close to the orbital plane, the stream from the vicinity of L_2 expands and is twisted under the action of the Coriolis force, resulting in the generation of a stellar wind that is typical of binary systems. We also believe that the thickness of this wind in the z direction grows with distance from the system. Thus, the secondary loses mass simultaneously through both libration points, L_1 and L_2, and, due to the flow from L_2, the mass transfer in the system is nonconservative.

To trace the generation of the binary wind via the stream from L_2, we carried out computations using a large spatial grid. The results of these computations are presented in Fig. 1c, which shows contours of equal density. We can see that the stream expands rapidly in the radial direction and generates a nearly homogeneous wind in the orbital plane. The velocity of the wind in the radial direction at the edge of the computational region is about 50–100 km/s, and the wind density is of the order of $10^{11}\mathrm{cm}^{-3}$.

3.1.2. Primary radius equal to 0.2 and $f_v = 0.1$. The contours of equal density and temperature and velocity field are shown in Fig. 3. The stream is now much more extended, since it moves in the larger Roche lobe of the primary. At the same time, the stream radius contracts by a factor of three to four compared to its radius at L_1 as it moves through this lobe, unlike the stream in the model of Lubow and Shu [10]. In this set of computations, the stream impacts the primary tangentially (the edge of the stream grazes the primary), resulting in the formation of a dense disk around the primary and of a less dense and prominent outer envelope surrounding the disk and both components of the system. As in the previous cases, at the outer boundary of the envelope, the gas motion is opposite to the rotation of the binary.

In this case, several shock waves form in the computational region. Shock I forms due to the

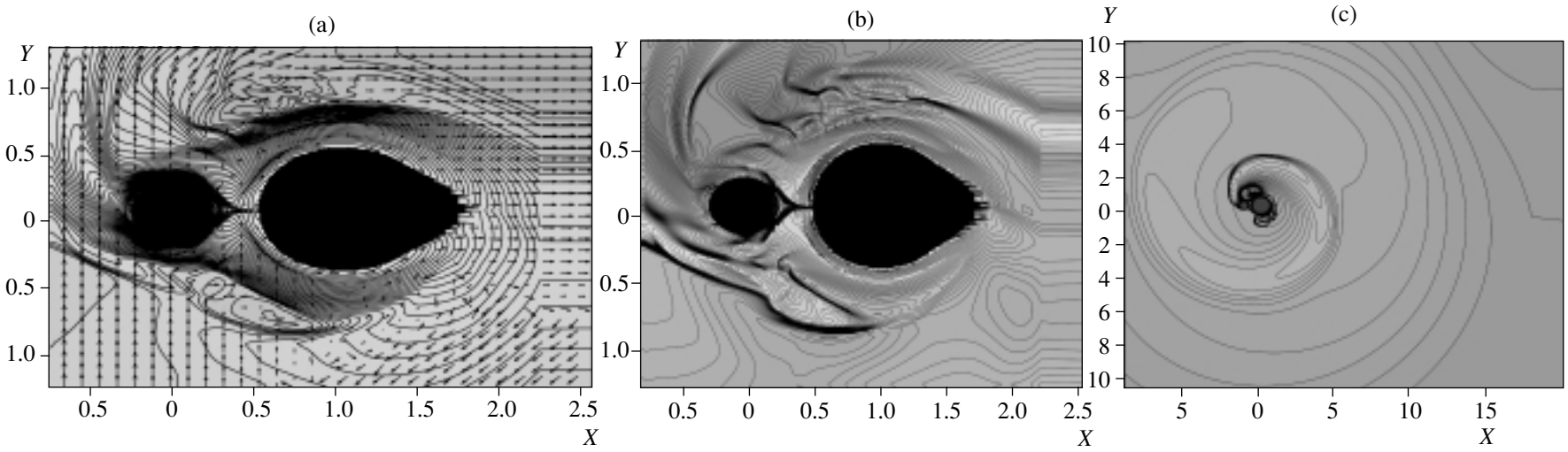

Fig. 1. (a) Contours of equal density and velocity vectors and (b) contours of equal temperature for the case $R_1 = 0.5$, $f_v = 0.5$, and $\dot{M} = 10^{-5} M_\odot/\mathrm{yr}$, with radiative cooling implicitly taken into account. Panel (c) shows the formation of a stellar wind from the binary system (obtained using a large computational grid).

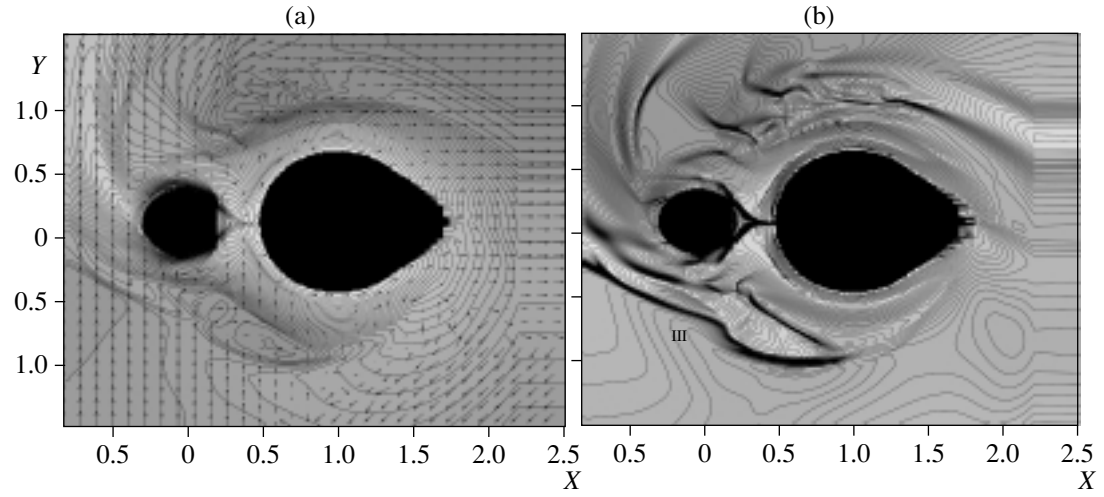

Fig. 2. (a) Contours of equal density and velocity vectors and (b) contours of equal temperature for $R_1 = 0.5$, $f_v = 0.1$, and $\dot{M} = 10^{-5} M_\odot/\mathrm{yr}$ with radiative cooling implicitly taken into account.

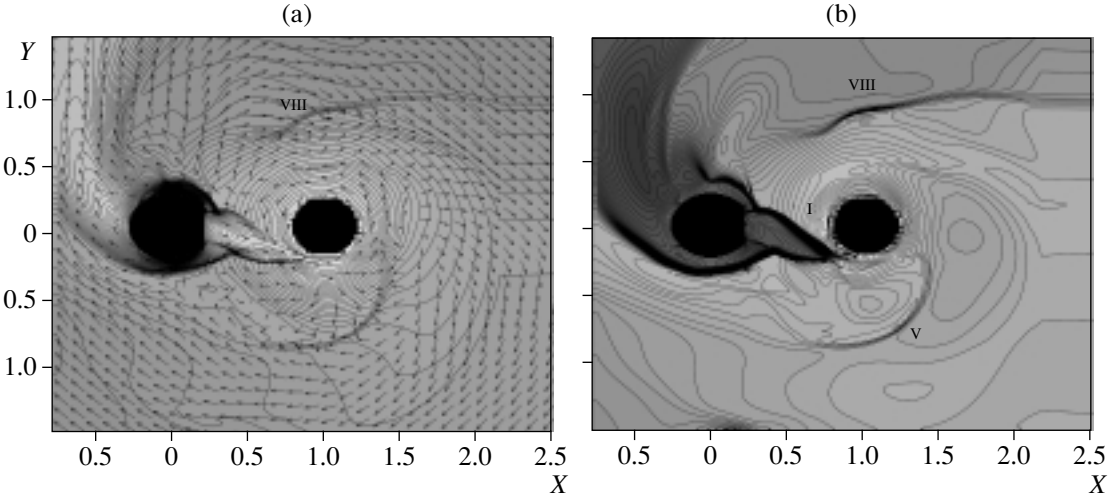

Fig. 3. Same as Fig. 2 for $R_1 = 0.2$ and $f_v = 0.1$.

interaction of the gas moving inside the Roche lobe of the primary and the upper edge of the stream. Shock V forms in the reverse flow of gas from L_3 to L_1. Shock VIII forms due to the interaction of the gas from the accretion disk with the gas from the outer envelope (close to orbital phase 0.25).

3.1.3. Primary radius equal to 0.1 and $f_v = 0.1$. Figure 4 shows the lines of equal density and temperature and the velocity field in the computational region. The radius of the stream at L_1 is 0.22. Due to the small radius of the primary, the stream does not impact the surface of this star, instead passing a short distance from it, of the order of the star's radius. The stream moving inside the primary Roche lobe forms a pronounced accretion disk (the number density in the disk is of the order of 10^{12}–10^{13} cm^{-3}), and some of the matter also flows into the circumstellar envelope, where the gas rotates opposite to the rotation of the binary system. The matter flowing from the stream into the circumstellar envelope interacts with the stream at its outer edge (at orbital phases near 0.75) and forms shock V. As in the case with $R_1 = 0.2$, shock I forms at the outer edge of the stream.

3.2. Models with Explicit Account of Radiative Cooling

3.2.1. Primary radius equal to 0.5 and $f_v = 0.5$ and 0.1. The contours of equal density and temperature and the velocity field are shown in Fig. 5

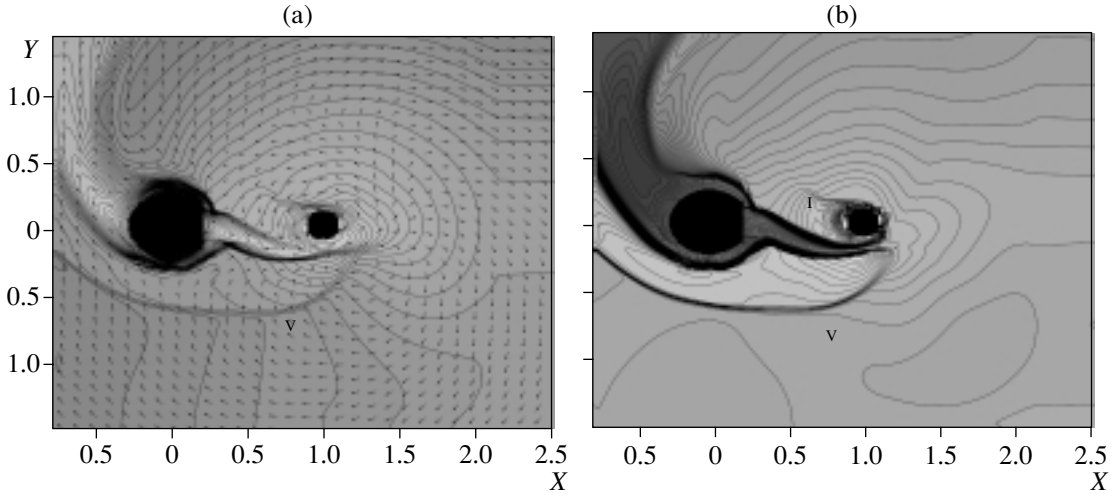

Fig. 4. Same as Fig. 2 for $R_1 = 0.1$ and f_v=0.1.

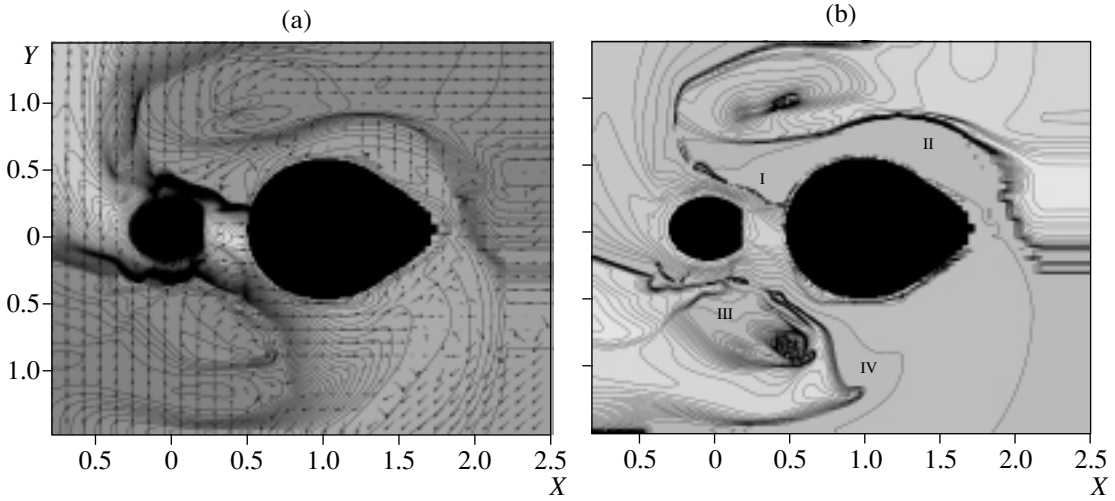

Fig. 5. Same as Fig. 2 for $R_1 = 0.5$ and $f_v = 0.5$ with radiative cooling explicitly taken into account.

($f_v = 0.5$) and Fig. 6 ($f_v = 0.1$). The radius of the stream in the vicinity of L_1 remains the same: 0.22. The explicit account of radiative cooling drastically changes the appearance of the envelopes in the close binary and makes their structure more complex. Only the asymmetry of the envelope for $f_v = 0.5$ and its symmetry for $f_v = 0.1$ remain. It is striking that the libration points L_4 and L_5 become more prominent. The minimum of the density and temperature in the envelopes is near these points. We can also see that the stream reflected from the surface of the primary interacts with the stream moving from L_1. As in the previous case, this results in the formation of a region of high temperature between L_1 and the surface of the primary. Computations of the mass transfer in an almost contact system with explicit account of radiative cooling show an appreciable increase in the temperature and density gradients in the envelopes, especially at their outer edges. The velocities in the outer envelope remain modest, and the envelopes are close to hydrostatic equilibrium. The stream of matter from L_2 is preserved.

We can see the system of shocks typical for the envelope of an almost contact binary in Figs. 5 and 6. Shock I propagates along the upper edge of the stream to the surface of the secondary and further to the right edge of the stream flowing from L_2. Shock II forms in the vicinity of L_4, close to the surface of the primary. This shock arises due to the collision of the gas that is rotating around the primary and the gas that is close to the outer border of the computational region and moves opposite to the rotation of the system; it resembles shock II in a close binary with a compact object. Also present in the system

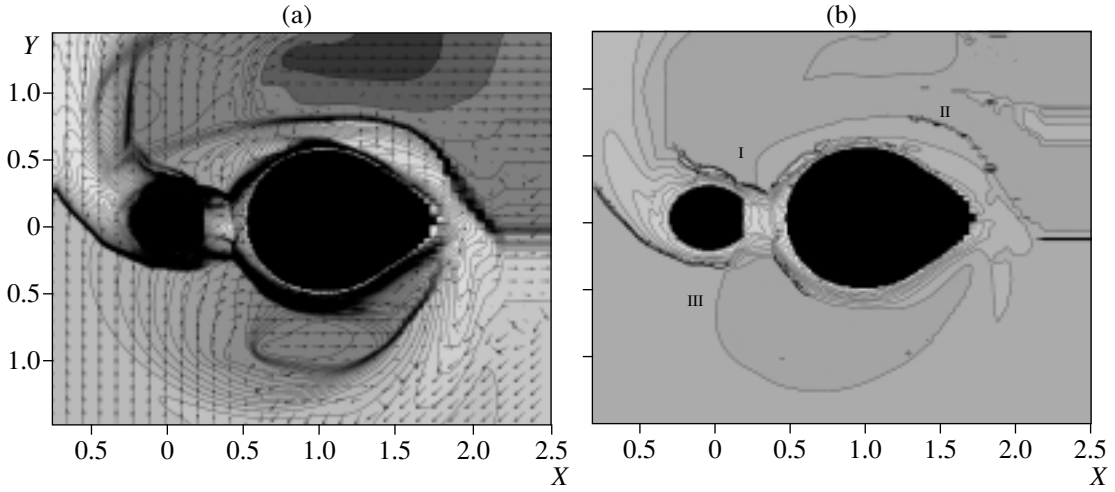

Fig. 6. Same as Fig. 5 for $R_1 = 0.5$ and $f_v = 0.1$.

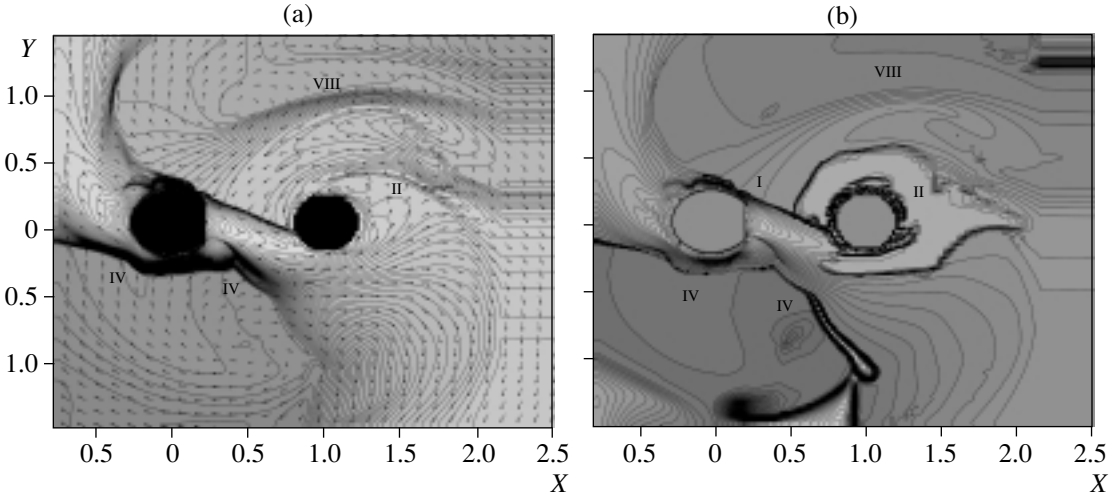

Fig. 7. Same as Fig. 5 for $R_1 = 0.2$ and $f_v = 0.1$.

are shock III, which forms below the line connecting the centers of the objects in the vicinity of L_5, and shock IV, which is located below the primary. Figure 6 shows shock V, which forms during the motion of the gas from L_3 to L_1.

3.2.2. Primary radius equal to 0.2 and $f_v = 0.1$. The lines of equal density and temperature and the velocity vectors are shown in Fig. 7. A more complex stream and envelope are formed in this case as well. First and foremost, the stream flowing near the primary expands, probably due to the decrease of the temperature in neighboring regions (the stream expands under the action of its own radial pressure gradient). Some fraction of the stream forms a compact, dense accretion disk, but a considerable part of the stream departs in the direction of orbital phases 0.7–0.8 and forms an envelope. Some of the stream and disk material flows out in the direction of L_3, then in the direction of L_1, generating shock IV, which has a very high temperature ($\sim$ 70 000–80 000 K). Such gas motions are apparently typical when radiative cooling is explicitly taken into account (it may be that an excess pressure from the disk gives rise to these outflows). In the rest of the space around the binary components, the temperature is $\sim$ 4000–18 000 K in the outer envelope (in very good agreement with the two-temperature model of Linnell [7]) and $\sim$ 30 000–100 000 K in the accretion disk. We can see in Fig. 7 that the spiral shock II appears (in the notation of Bisikalo *et al.* [18]). This wave apparently forms when the matter from the disk interacts with the matter of the circumstellar envelope. The parameters of this shock are maximum

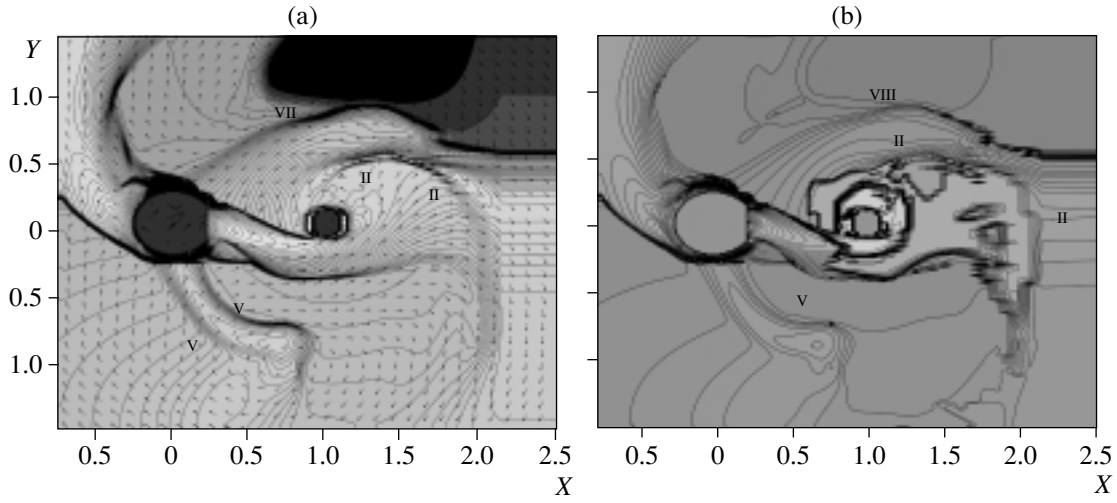

Fig. 8. Same as Fig. 5 for $R_1 = 0.1$ and $f_v = 0.1$.

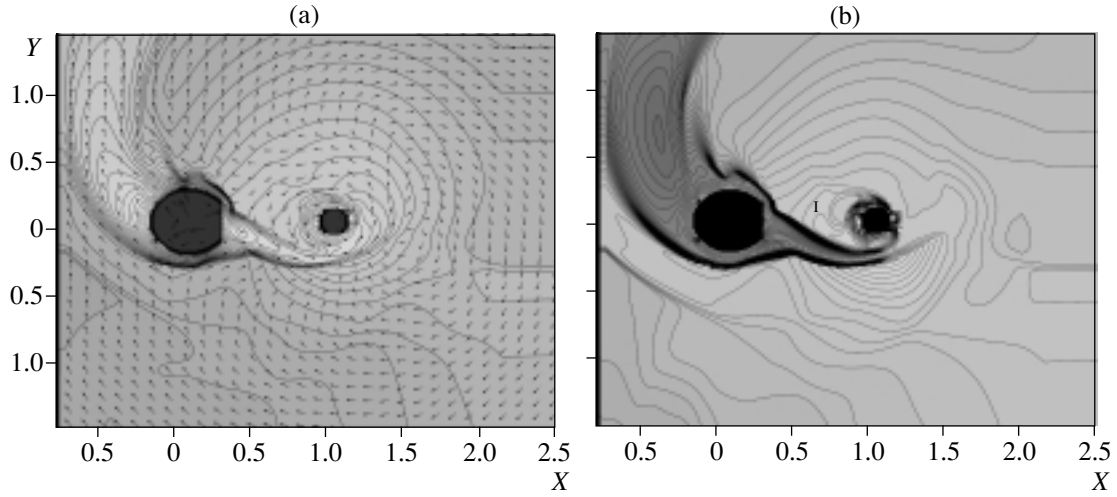

Fig. 9. Same as Fig. 2 for $R_1 = 0.1$, $f_v = 0.1$, and $\dot{M} = 4\times10^{-5} M_\odot/\mathrm{yr}$.

at the outer boundary of the computational region, where the interaction is most intense. Shock VIII bounds the accretion disk from the outside, and forms due to interaction of the disk gas with the gas of the outer envelope.

3.2.3. Primary radius equal to 0.21 and $f_v = 0.1$. The lines of equal density and temperature and the velocity field are shown in Fig. 8. As for the computations for other primary radii, a large number of shocks form in this case. Figure 8 shows that the stream moving inside the primary Roche lobe also flows close to the primary, as in the case of implicit account of radiative cooling, but this stream now expands considerably, its edge slightly grazing the surface of the star. The origin of this expansion is the radial pressure gradient.

Figure 8 also shows that the stream from L_1 is bounded on two sides by shocks (shock I in [18, 19]). As the stream spreads, the matter that initially moved toward L_3 turns toward the direction of L_1, moving opposite to the rotation of the system. When this matter encounters the gas of the outer envelope, shock V forms, which turns toward the direction of orbital phase 0.9 under the action of the Coriolis force. The stream expanding near the primary spreads into a broad range of orbital phases from 0.25 to 0.6, and forms a large, dense accretion disk and a dense outer envelope (the number density in this envelope is of the order of 10^{11}–10^{12} cm^{-3}, while that in the disk is $\sim 10^{12}$–10^{13} cm^{-3}). At phases close to 0.25, the accretion disk is bounded by shock VIII, which forms when the gas of the disk interacts with the gas of the outer envelope moving in the opposite direction. With the given parameters of the stream, we can see a very

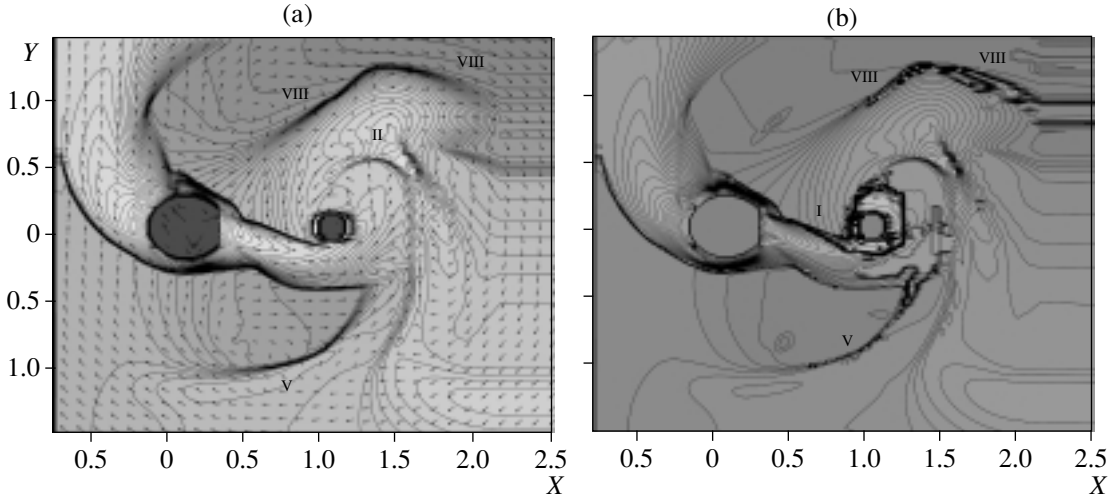

Fig. 10. Same as Fig. 9 with radiative cooling explicitly taken into account.

strong spiral shock II, which is much more prominent than in the case $R_1 = 0.2$. This shock has a complex structure and, originating in the disk, extends to the region of L_3. There is a stream from the vicinity of L_2 (it is still nonconservative), which is bounded by the shock that forms due to the interaction of the stream gas with the gas of the outer envelope.

4. MODELS FOR A MASS-TRANSFER RATE OF $4 \times 10^{-5} M_\odot$/yr WITH EXPLICIT AND IMPLICIT ACCOUNT OF RADIATIVE COOLING ($R_1 = 0.1$)

If radiative cooling is taken into account implicitly, the radius of the stream for a mass-transfer rate of $\dot{M} = 4{\times}10^{-5} M_\odot$/yr is 0.29. The contours of equal temperature and density and velocity field at the end of the computations are shown in Fig. 9 (implicit radiative cooling) and Fig. 10 (explicit radiative cooling). Figure 9 shows that the powerful stream moving inside the Roche lobe of the primary considerably (by a factor of two to three) reduces its radius and, without passing close to the primary as it does in the case of a lower mass-transfer rate, forms a dense accretion disk around the primary. This dense disk (10^{13}–10^{14} cm^{-3}) moves the stream even further from the primary, to a distance comparable to the radius of the star. Thus, the stream moving in the Roche lobe of the primary forms a very powerful disk whose radius is approximately equal to the radius of the lobe and an outer envelope whose density is of the order of 10^{12}–10^{13} cm^{-3}. The interaction of the stream with the accretion disk occurs without the formation of a hot spot, but with the formation of a hot line; i.e., the stream does not impact the disk, but, on the contrary, the accretion disk impacts the edge of the stream.

Figure 9 shows that no prominent system of shocks forms when radiative cooling is taken into account explicitly. This is apparently due to the specific properties of the stream's motion inside the Roche lobe of the primary in this case. We can see only the rather weak shock I at the upper edge of the stream, close to the primary, and another shock that bounds the stream from L_2; as in all the previous cases, this latter stream makes the mass transfer in the system nonconservative.

If radiative cooling is taken into account explicitly with the same parameters for the system, many more shocks form, even compared to models with the same parameters but a lower mass-transfer rate (Fig. 10). First and foremost, the stream moving inside the Roche lobe of the primary expands and grazes the surface of the star. The origin of this expansion is the same—the action of the radial pressure gradient. Figure 10 shows that the stream from L_1 is bounded on two sides by shocks (shock I). The matter that spreads from the stream also forms a shock (shock V), which is deflected in the direction of orbital phase 0.9 under the action of the Coriolis force. Figure 10 also shows that the stream that expands close to the primary spreads in the direction of orbital phases from 0.25 to 0.75, and forms a large accretion disk (which nearly fills the primary Roche lobe) with a number density of close to 10^{13}–10^{14} cm^{-3} and a dense outer envelope with a number density of the order of 10^{12} cm^{-3}. The accretion disk is bounded at phases near 0.25 by shock VIII, which originates due to the interaction of the disk and outer envelope, which are moving in opposite directions. With the given parameters of the system, we can see a very strong spiral shock II, which is more prominent than in the case $R_1 = 0.2$. This shock has a complex

structure and, originating in the disk, extends to the vicinity of L_3.

Figures 9 and 10 show that there is a stream of matter from the vicinity of L_2, which is bounded by the shock that forms due to the interaction of the stream with the outer envelope. As in the previous cases, this stream from L_2 makes the mass transfer in the system nonconservative.

5. MODELS TAKING INTO ACCOUNT THE STELLAR WIND OF THE SECONDARY

In this set of computations, we took into account the stellar wind from the secondary. Based on the component masses, it would seem more appropriate in the first instance to take into account the stellar wind of the primary. However, due to the presence of the accretion disk in the system, this wind would probably flow only above and below the orbital plane, since the accretion disk will hinder the motion of the wind in the orbital plane. Hence, this wind will have an effect only in three-dimensional computations. On the other hand, nothing prevents the formation of a stellar wind from the secondary and its initial outflow in the orbital plane. Thus, such a wind can influence the results of two-dimensional computations.

We took the wind velocity to correspond to the velocity in the model of Barlow and Cohen [20], $\dot{M} = 5\times10^{-13}L^{1.2}M_\odot$/yr, expected for a star with a mass of 2 $M_\odot$. The wind velocity was determined as follows. We first determined the velocity of the wind from the secondary using the asymptotic formula and the parameters of the system. We then derived the density at the outer boundary of the model atmosphere corresponding to the secondary and the density at the position in the wind where the velocity attains the value V_∞. Using this density and V_∞, we could then find the velocity of the wind of the secondary. In our computations, the wind velocity is about 80 km/s at a distance of 1.2 (1.0 corresponds to the mean radius of the secondary Roche lobe); this corresponds to the very weak wind of such a moderate-mass star. The mass-transfer rate in this set of computations is also $4 \times 10^{-5}M\odot$/yr.

We continued our computations over 17 orbital periods to ensure that a steady-state was attained in the computational region. The contours of equal density and temperature and the velocity field at the end of the computations are shown in Fig. 11 (with radiative cooling taken into account implicitly) and Fig. 12 (with radiative cooling taken into account explicitly). The initial radius of the stream is 0.28 (the change of the radius compared to the previous set of computations is due to the effect of the stellar wind on the formation of the stream: this wind slightly suppresses the formation of the stream at L_1).

As Figs. 11 and 12 show, with such a high mass-transfer rate, the stellar wind only weakly influences the structure of the stream at L_1 and its motion inside the Roche lobe of the primary. The only appreciable difference is that, due to the action of the stellar wind in the vicinity of L_1, the stream retains its overall structure, but is oriented along a larger angle to the line of the component centers. This results in a small shift in the direction of the stream as it moves inside the Roche lobe of the primary. As in the previous case, the stream forms a powerful accretion disk and circumstellar envelope. Figure 11b very clearly shows shock I (the hot line) at the edge of the stream and shocks III and IV, which are parallel to the surface of the secondary in the region where the donor wind is formed. Figure 11 shows that the stellar wind immediately manifests itself via the appearance of a system of shocks that are generated by the interaction of the wind with the envelopes of the binary system. We label these shocks with Roman numerals starting from X, to distinguish the shocks due to the interaction of the stellar wind with the envelopes of the binary system.

As Fig. 11 shows, if radiative cooling is taken into account implicitly, the wind is confined on all sides by the accretion disk, outer envelope, and stream from L_2, and moves in this restricted area. The interaction of the stellar wind with the accretion disk results in the formation of shock X, which bounds the wind from the side of the accretion disk. The interaction of the stellar wind with the stream from the vicinity of L_2 results in the formation of shock XI. The interaction of the wind with the outer envelope (near orbital phases 0.6–0.9) generates shock XII. Thus, three shocks are generated by the interaction of the wind with the gas flows in the binary system. In addition, Fig. 11b shows shocks III and IV in the region where the stellar wind flows from the atmosphere of the secondary.

The motion of the wind and its interaction with the envelopes changes markedly when radiative cooling is taken into account explicitly. In this case, the wind from the secondary is not confined by the envelopes (since the temperature in them is considerably lower), and the wind dominates in the space around the binary components. The velocity of the wind in the region where its formation is completed remains close to 80 km/s, but it now fills almost the entire space between the binary components.

Figure 12 shows that the formation of the stream is not substantially changed compared to the previous sets of computations. However, the stream is appreciably compressed by the wind in the vicinity of L_1, and its motion inside the Roche lobe of the primary differs from the case with explicit account for radiative cooling. The stream splits near the primary, with the bulk of the stream moving in the direction of orbital phases 0.65–0.85, while, after moving around

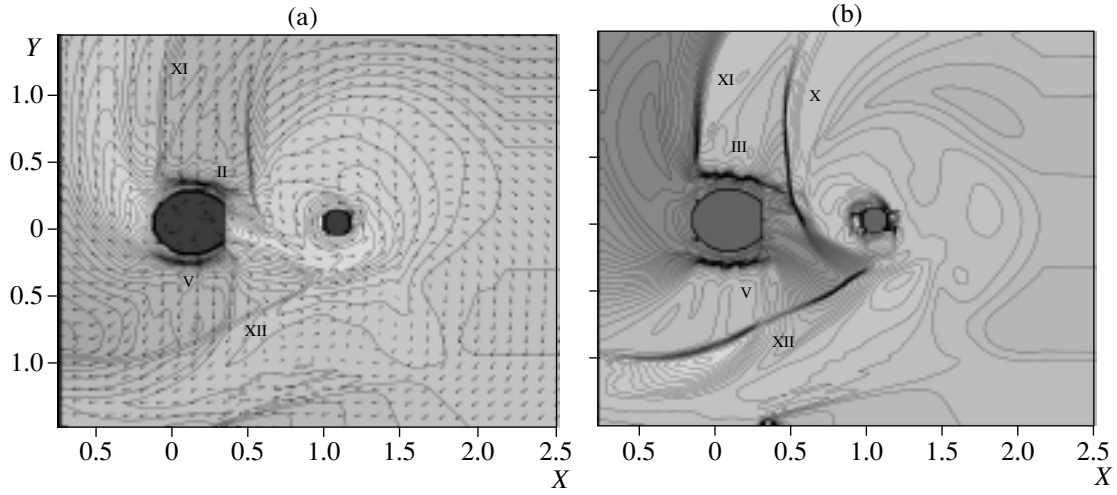

Fig. 11. Same as Fig. 9 including the stellar wind of the secondary.

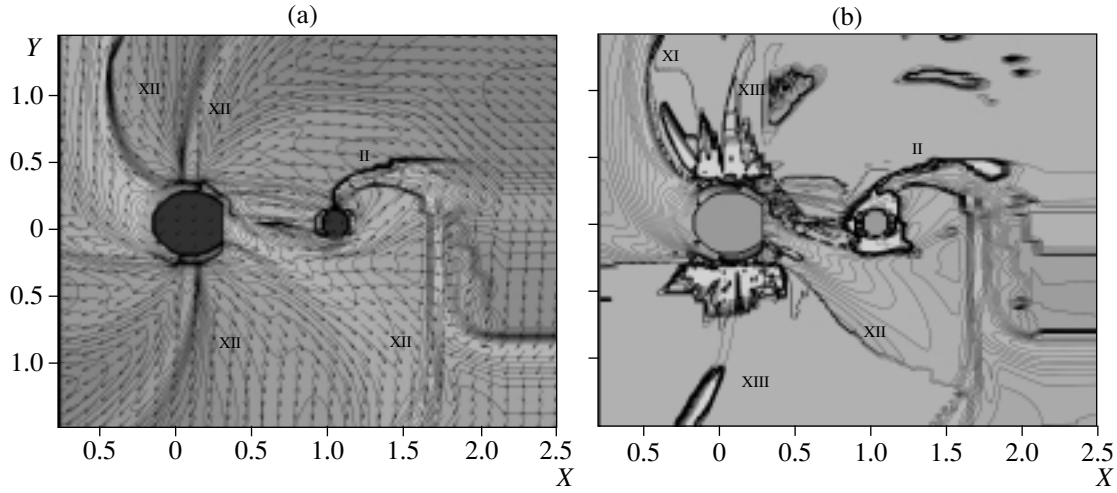

Fig. 12. Same as Fig. 9 with explicit account of radiative cooling and including the stellar wind of the secondary.

the primary, the remainder interacts with the stellar wind flowing above the line of component centers. The system of shocks changes. Most importantly, the new shock XIII appears, which originates in the wind itself and is due to the interaction of different parts of the wind under the action of the Coriolis force. This shock forms both below and above the line of component centers, and the Coriolis force curves its motion in the direction opposite to the rotation of the binary. Shock XI has the same position as in the case with implicit account of radiative cooling. Shock XII, which originates due to the interaction of the wind with the stream (its lower edge), is very indistinct and stretched.

As the velocity vectors in Fig. 12 clearly show, there is virtually no accretion disk around the primary. It is essentially swept out by the stellar wind, and we can see that the spiral shock II in this version of the cooling is formed by the interaction of the wind with the stream, which moves around the primary and ends up above the line of component centers, in the region of orbital phases 0.3–0.4. Thus, as in the computations with explicit account of radiative cooling and without stellar wind, the spiral shock II forms; but, in this case, when there is no accretion disk, this shock is due to the interaction of the wind and the stream from L_1.

6. DISCUSSION

We have carried out two-dimensional hydrodynamical computations of the mass-transfer in the β Lyr binary system for various sets of parameters of the system, both explicit and implicit account of radiative cooling, various assumptions about the character of the interaction of the stream and the primary,

Results of model computations of the physical characteristics of the β Lyr system

Properties of gaseous structures	Implicit account of radiative cooling						Explicit account of radiative cooling					
	$\dot{M} = 10^{-5} M_\odot$/yr				$\dot{M} = 4\times10^{-5} M_\odot$/yr		$\dot{M} = 10^{-5} M_\odot$/yr				$\dot{M} = 4\times10^{-5} M_\odot$/yr	
	$R_1 = 0.5$ $f_v = 0.5$	$R_1 = 0.5$ $f_v = 0.1$	$R_1 = 0.2$ $f_v = 0.1$	$R_1 = 0.1$ $f_v = 0.1$	$R_1 = 0.1$ $f_v = 0.1$	$R_1 = 0.1$ $f_v = 0.1$ +stellar wind	$R_1 = 0.5$ $f_v = 0.5$	$R_1 = 0.5$ $f_v = 0.1$	$R_1 = 0.2$ $f_v = 0.1$	$R_1 = 0.1$ $f_v = 0.1$	$R_1 = 0.1$ $f_v = 0.1$	$R_1 = 0.1$ $f_v = 0.1$ +stellar wind
Radius of the stream at L_1	0.22				0.29	0.28	0.22				0.29	0.28
Stream in the Roche lobe of the primary	Expands		Narrows by factor 3–4	Narrows	Narrows by factor 2–3	Expands	Expands				Expands	
Interaction of the stream with the primary	Central impact		Tangential impact	No impact	No impact	No impact	Central impact		Off-center impact	Tangential impact	Tangential impact	No impact
Accretion disk	Absent	Absent	Dense disk	Large dense disk 10^{12}–10^{13} cm^{-3}	Large, dense disk 10^{13}–10^{14} cm^{-3}	Dense disk	Absent	Absent	Dense disk 10^{12}–10^{13} cm^{-3}	Large dense disk 10^{13}–10^{14} cm^{-3}	Large, dense disk	Absent
Gaseous envelope	Asymmetric	Symmetric	Extended		Dense 10^{12}–10^{13} cm^{-3}	Dense extended	Asymmetric	Asymmetric	Extended	Dense 10^{11}–10^{12} cm^{-3}	Dense 10^{12}–10^{13} cm^{-3}	Dense extended
Shock waves	Absent	III	I, V, VIII	I, V	I	I, III, IV, X, XI, XII	I, II, III, IV	I, II, III, V	I, II, IV, VIII	I, II, V, VIII	I, II, V, VIII	II, XI, XII, XIII
Stream at L_1	Always present											

and with and without including the stellar wind from the secondary. Thus, our computations represent a study of the influence of various physical processes and adopted parameters of a close binary system on the mass transfer in the system. The main results are summarized in the Table, which shows that the structure of the envelopes that form as a result of the mass transfer depends very strongly on the adopted model assumptions. First and foremost, the results depend on the radius of the primary.

When the primary radius is large, $R_1 = 0.5$, the stream from L_1 flows almost without any deflection, and therefore impacts the center of the star. As a result, only a circumstellar envelope is formed, without an accretion disk. The formation of an envelope when $R_1 = 0.5$ depends strongly on the coefficient f_v: when $f_v = 0.5$, an asymmetric circumstellar envelope forms, while the envelope is symmetric relative to various orbital phases when $f_v = 0.1$. For this radius, shocks essentially do not form if radiative cooling is taken into account implicitly, while four shock waves form when cooling is taken into account explicitly.

When the radius of the primary is $R_1 = 0.2$, the stream's interaction with the star is tangential when radiative cooling is taken into account implicitly, while the stream impacts the edge of the star in the case of explicit account of radiative cooling. As a result, a compact, dense disk and extended envelope with a complex system of shocks form in both cases.

When the radius of the primary is $R_1 = 0.1$, when radiative cooling is taken into account implicitly, the stream does not interact with the star, and forms a dense disk around the star, comparable in size to the primary Roche lobe, and an extended, dense envelope with a small number of shocks. If radiative cooling is taken into account explicitly, the stream's interaction with the star is tangential, and a large, dense accretion disk and extended, dense envelope with a complex system of shocks are formed. The dense accretion disk deflects the stream from the line of component centers even more, and the interaction of the stream with the disk produces a hot line instead of a hot spot.

In all cases, the secondary also loses mass through L_2. The stream from L_2 leaves the system, creating a nearly homogeneous wind in the orbital plane. The amount of matter lost from L_2 is close to 40% of the mass flowing through L_1. When radiative cooling is taken into account explicitly, the temperature in the envelope is of the order of 4000–18 000 K, close to values derived from observations [7]. If the stellar wind from the secondary is taken into account, new shocks form. When radiative cooling is taken into account explicitly, the stellar wind disrupts the accretion disk. However, this result may be due to to the two-dimensional nature of our computations.

It is likely that, in three-dimensional computations, the wind that is reflected from the accretion disk interacts with the wind from the primary above and below the orbital plane. It is quite possible that the jets observed in the system [5] are due to both the interaction of the stream and accretion disk, and interactions between the stellar winds in the system. Thus, the presence of stellar winds changes significantly the pattern of the gaseous flows in the β Lyr system, making it much more complex. It is clear that these winds must be taken into account if we wish to obtain a realistic model of this system.

Considering all our computational results, we conclude that the most realistic version of our computations for the β Lyr system is the one that includes the stellar wind of the secondary, takes into account radiative cooling explicitly, and has a primary radius equal to 0.1.

Bearing in mind the large radius of the stream in the vicinity of L_1, it would be worthwhile in the future to carry out three-dimensional computations of the mass transfer in this system based on the results of these two-dimensional computations.

7. CONCLUSIONS

Our analysis of the results of our computations led us to the following conclusions.

(1) The radius of the stream in the vicinity of L_1 is about 0.22 for a mass-transfer rate of $10^{-5} M\odot$/yr and about 0.29 for a mass-transfer rate of $4\times10^{-5} M\odot$/yr.

(2) The motion of the stream in the Roche lobe of the primary leads to the formation of a large accretion disk with a number density of 10^{13}–10^{14} cm^{-3}. Instead of a hot spot on the disk, a hot line forms at the edge of the stream, together with a circum-binary envelope with a number density of 10^{12}–10^{13} cm^{-3}. Thus, instead of the stream impacting the disk, we have a situation in which the disk impacts the stream.

(3) When radiative cooling of the moving gas is taken into account explicitly, the stream expands near the primary, in contrast to the predictions of the model of Lubow and Shu [10].

(4) The pattern of the interaction of the stream and accretion disk (or primary) depends, apart from the radius of the primary, on the rate of mass transfer through L_1. In the case of high mass-transfer rates (of the order of $4.0\times10^{-5} M\odot$/yr), the denser accretion disk moves the stream away from the primary.

(5) The system of shocks that forms depends on the treatment of radiative cooling. Shock I (the hot line), which forms at the edge of the stream facing the primary (in the direction of orbital phases 0.2–0.4) is

present in all models, while the spiral shock II forms in all models with explicit account of radiative cooling.

(6) In all the computations, the secondary loses mass through two libration points—L_1 and L_2. The flow through L_1 is completely conservative, while the flow through L_2 is completely nonconservative. This results in the outflow from the system of about 30% of the matter lost by the secondary.

(7) When the stellar wind from the secondary is taken into account, the shocks associated with the interaction of the wind and envelopes that form due to the outflow of matter through L_1 and L_2 dominate in the system, with the exception of the spiral shock II, which is present in models both with and without this stellar wind.

REFERENCES

1. I. Hubeny and M. J. Plavec, Astron. J. **102**, 1156 (1991).
2. R. E. Wilson, Astrophys. J. **189**, 319 (1974).
3. R. E. Wilson and D. Terrell, Ann. N.Y. Acad. Sci. **675**, 65 (1992).
4. I. Hubeny, P. Harmanec, and S. N. Shore, Astron. Astrophys. **289**, 411 (1994).
5. P. Harmanec, F. Morand, and D. Bonneua, Astron. Astrophys. **312**, 877 (1996).
6. A. P. Linnell, I. Hubeny, and P. Harmanec, Astrophys. J. **509**, 379 (1998).
7. A. P. Linnell, Mon. Not. R. Astron. Soc. **319**, 255 (2000).
8. J. P. De Creve and A. P. Linnell, Astron. Astrophys. **291**, 786 (1994).
9. D. V. Bisikalo, P. Harmanec, A. A. Boyarchuk, *et al.*, Astron. Astrophys. **353**, 1009 (2000).
10. S. Yh. Lubow and F. H. Shu, Astrophys. J. **198**, 383 (1975).
11. R. L. Kurucz, Astrophys. J., Suppl. Ser. **40**, 1 (1979).
12. V. V. Nazarenko, L. V. Glazunov, and V. G. Karetnikov, Astron. Zh. **78**, 525 (2001) [Astron. Rep. **45**, 452 (2001)].
13. D. Molteni, G. Belvedere, and G. Lanzafame, Mon. Not. R. Astron. Soc. **249**, 748 (1991).
14. D. P. Cox and E. Daltabuilt, Astrophys. J. **167**, 113 (1971).
15. J. M. Blondin, M. T. Richards, and M. L. Malinowski, Astrophys. J. **445**, 939 (1995).
16. P. Harmanec, Astron. Astrophys. **266**, 307 (1992).
17. O. M. Belotserkovskiĭ and Yu. M. Davydov, *Large Particles in Gas Dynamics* (Nauka, Moscow, 1982) [in Russian].
18. D. V. Bisikalo, A. A. Boyarchuk, V. M. Chechetkin, *et al.*, Mon. Not. R. Astron. Soc. **300**, 39 (1998).
19. D. V. Bisikalo, A. A. Boyarchuk, V. M. Chechetkin, *et al.,* Astron. Zh. **43**, 797 (1999).
20. M. J. Barlow and M. Cohen, Astrophys. J. **213**, 737 (1977).

Translated by L. Yungel'son

Astronomy Reports, Vol. 47, No. 12, 2003, pp. 1027–1037. Translated from Astronomicheskiĭ Zhurnal, Vol. 80, No. 12, 2003, pp. 1114–1126.
Original Russian Text Copyright © 2003 by Nazarenko, Glazunova.

Two-Dimensional Hydrodynamical Modeling of Mass Transfer in Semidetached Binaries with Asynchronously Rotating Components

V. V. Nazarenko and L. V. Glazunova

Astronomical Observatory, Odessa National University, Odessa, Ukraine

Received December 15, 2002; in final form, July 26, 2003

Abstract—We have modeled the mass transfer in the three semidetached binaries U Cep, RZ Sct, and V373 Cas taking into account radiative cooling both implicitly and explicitly. The systems have asynchronously rotating components and high mass-transfer rates of the order of $10^{-6}M_{\odot}$/yr; they are undergoing various stages of their evolution. An accreting star rotates asynchronously if added angular momentum is redistributed over the entire star over a time that exceeds the synchronization time. Calculations have indicated that, in the model considered, mass transfer through the point L_1 is unable to desynchronize the donor star. The formation of an accretion disk and outer envelope depends on the component-mass ratio of the binary. If this ratio is of the order of unity, the flow makes a direct impact with the atmosphere of the accreting star, resulting in the formation of a small accretion disk and a relatively dense outer envelope. This is true of the disks in U Cep and V373 Cas. When the component-mass ratio substantially exceeds unity (the case in RZ Sct), the flow forms a large, dense accretion disk and less dense outer envelope. Taking into account radiative cooling both implicitly and explicitly, we show that a series of shocks forms in the envelopes of these systems. © *2003 MAIK "Nauka/Interperiodica"*.

1. INTRODUCTION

At certain stages in the evolution of close binary systems, one of the stars fills its Roche lobe and starts losing matter through the vicinity of the inner Lagrange point (L_1). This matter carries away some of the angular momentum of the outflowing star (donor), while the accretion of this matter onto the secondary (accretor) adds angular momentum to this star. Due to this, we should observe desynchronization of the rotation and revolution of the components in some close binaries at the end of or immediately after the rapid mass-transfer phase of their evolution [1].

The modern theory of tidal interactions in detached binaries with main-sequence components is consistent with the observed parameters of many systems [2]. For components with both radiative and convective envelopes, the theory provides formulas that can be used to estimate the synchronization and circularization time, or to determine a lower limit for this time for a specific binary. On this basis, it was concluded that the components of essentially all semidetached binaries should be synchronized, and their orbits circular. Since the semidetached phase lasts 10^8 years in Algols, while their average synchronization time is about 10^5 years, the number of systems with asynchronously rotating components should be slightly more than one thousandth of all Algols. However, over the last ten years, the rotation of the accreting components in more than 20 of several hundred well-studied Algol-type systems have been found to be substantially asynchronous [3]. Some semidetached systems with massive components also display asynchronous rotation. This number of Algol-like systems with asynchronous rotation cannot be explained purely by suggesting that these systems have not yet become synchronized after the rapid mass-transfer phase.

It is important to study Algol-like systems with asynchronously rotating components in order to understand how to correctly calculate the transfer of mass and angular momentum in the theory of binary evolution. We have modeled the mass transfer in several semidetached systems that have the maximum mass-transfer rates observed for Algol-like systems. It is possible that the current mass-transfer rates are sufficient to maintain asynchronicity due to accretion or mass loss through L_1 in some semidetached systems with asynchronously rotating components. To test this hypothesis, we carried out two-dimensional hydrodynamical modeling of the mass transfer in the orbital planes of three semidetached systems: RZ Sct, U Cep, and V373 Cas. These systems have asynchronously rotating components, and their high rates of mass transfer through the inner Lagrange point could potentially bring about the observed asynchronicity.

Table 1 contains the basic parameters of the systems. The first and second columns present the studied systems and spectral types of their components.

1063-7729/03/4712-1027$24.00 © 2003 MAIK "Nauka/Interperiodica"

Table 1. Basic parameters of the studied systems

System	Spectral type	F	$\dot{M}$, $M_\odot$/yr	P, days	M_{acc}, $M_\odot$	q	R_{acc} ($A = 1$)
V373 Cas	B1(II–Ib)+0.5III	2	10^{-5}–10^{-6}	13.4	14	0.74	0.15
U Cep	B7V+G8IV	8	10^{-6}–10^{-7}	2.49	3.6	1.49	0.18
RZ Sct	B1Ib+A2IV	4	10^{-6}	15.2	16.3	5.26	0.17

The third column contains the asynchronicity factor for the accretor (for V373 Cas, for both the accretor and donor); i.e., the ratio of the rotational and orbital velocities of the star. The fourth column presents the observed mass-transfer rates, the fifth column the rotational period of the system, the sixth and seventh columns the mass of the donor and the component-mass ratio (accretor to donor), and the last column the radius of the accretor. The observed characteristics of the systems are as follows.

373 Cyg is a massive eclipsing B1(II-Ib) + B0.5III binary with an orbital eccentricity of 0.13. The system displays variable H_α emission, providing evidence for instability of the mass transfer from the donor. The observation of polarization from the system is ascribed to the existence of a aspherical common envelope. Evolutionary calculations yield a lifetime for the system of the order of $(7–8) \times 10^6$ yrs [4]. Both components rotate asynchronously, with the rotational velocity being twice the orbital velocity. The rotational velocities of the components have been determined from observations of the HeI $\lambda\lambda$4388, 4471, 4922 Å lines. The more massive star fills its Roche lobe and loses mass through L_1, suggesting that V373 Cas has just started the first mass-transfer phase (case A). To explain the asynchronicity of the rotation of the components, Lyubimkov *et al.* [4] suggested that the donor has not yet become synchronized (i.e., we observe initial asynchronicity), while the accretor has gained angular momentum due to the accretion.

RZ Sct is a massive long-period B1Ib + A2IV system. Its evolutionary status is similar to that of Algol-like binaries; i.e., the system has undergone the first mass-transfer phase, and the mass transfer is currently proceeding on the nuclear timescale. The system displays strong two-peaked Hα emission, confirming the presence of a disk envelope around the accretor [5]. The angular velocity of rotation of the accretor exceeds that of its orbital motion by a factor of four. The rotational velocities determined from different spectral lines differ appreciably (183 km/s from HeI λ4471 Å versus 282 km/s from MgII λ4481 Å [6]). The descending and ascending arms of the primary minimum of the light curve display an asymmetry due to absorption in the gas flow. The estimated density in the flow is 10^{13} cm^{-3}, with the mass-transfer rate being about $6 \times 10^{-7} M_\odot$/yr [7]. Olson and Etzel [6] proposed that the observed asynchronicity of the rotation of the accretor is due to a high mass-transfer rate in the system. The rate of change of the system's period is $\Delta P/P = 2 \times 10^{-6}$ [8].

The eclipsing binary U Cep is classified as a typical Algol with B7V + G8IV components. However, the estimated mass-transfer rate in this system is of the order of $10^{-6} M_\odot$/yr, whereas the standard value for Algols is 10^{-8}–$10^{-9} M_\odot$/yr. The high mass-transfer rate is the origin of all the photometric and spectral peculiarities of this system—distortions of the light curve and radial-velocity curve, the presence of radio and X-ray radiation, and variable Hα emission—which indicate the presence of flows and an inhomogeneous accretion disk in the system. The rotational angular velocity of the accretor exceeds its orbital velocity by a factor of eight [9]. Burnett and Etzel [10] suggest that the asynchronicity of the rotation and revolution of the primary results from ongoing accretion onto the star. The relative rate of change of the system's period $\Delta P/P$ is of the order of 10^{-5}; quasi-periodic variations of the period with a period of about 12 years have also been noted [11].

2. HYDRODYNAMICAL MODELING OF MASS TRANSFER IN CLOSE BINARIES

The mass-loss rates due to the stellar winds of the components of these three systems do not exceed $10^{-7} M_\odot$/yr, while the estimated mass-transfer rate through L_1 is of the order of $10^{-6} M_\odot$/yr. Therefore, we will neglect the loss of mass and angular momentum due to stellar wind. The degree of Roche-lobe overfilling by the donor is determined by requiring that the observed and calculated mass-transfer rates through L_1 be equal. Our approach to the modeling of mass transfer in close binary systems is based on two points:

(1) we derive the flow structure in the vicinity of L_1 by modeling the formation of the flow in this region;

(2) the initial structure of the donor atmosphere in the vicinity of L_1 is specified by a Kurucz model atmosphere [12], which makes it possible to take into account the properties of the mass transfer in a

particular close binary, since the atlas [12] covers a large range of atmospheric effective temperatures and gravitational accelerations.

It is also important to adopt an approach of this type because, according to the calculations of Lubow and Shu [13], the flow structure in the radial direction remains virtually constant as the flow moves toward the accretor. Consequently, the initial structure of the flow will be important when calculating the interactions between the flow and the atmosphere of the accretor. The use of our approach may also become important for studies of systems in which the initial flow structure differs dramatically from the model of Lubow and Shu [13] (for example, for long-period binaries).

To model the mass transfer in the close binaries, we solved the nonsteady-state Euler equations using the "large particles" method developed by Belotserkovskiĭ and Davydov [14]. This method was modified by introducing artificial viscosity that was second order in space and time in the first stage of the calculations. In practice, each time step took around 2–3 s, and a steady state was reached after 300 000–500 000 steps. A detailed description of the calculation technique is given in [15, 16].

Two-dimensional hydrodynamical modeling of mass transfer in a close binary containing normal stars without accurate calculation of the formation of the flow through L_1 was also carried out by Blondin *et al.* [17] for β Per (taking cooling into account) and by Richards and Ratliff [18] for β Per and TT Hyd. Three-dimensional hydrodynamical modeling of the mass transfer in β Lyr was performed by Bisikalo *et al.* [19].

3. INITIAL MODEL AND THE BINARY MODEL. BOUNDARY CONDITIONS

We used a standard model for the binary: the rotation of the components about their axes was assumed to be synchronized with their orbital revolution, the gravitation of the components was treated in a Roche-lobe approximation, all rotational effects (centrifugal and Coriolis forces) were taken into account, and the orbits of the components were assumed to be circular. For simplicity, we also assumed the gas to consist purely of hydrogen. We used both implicit and explicit methods to take into account radiative cooling in the moving gas. In the implicit approach, the adiabatic index in the equation of state for an ideal gas is decreased to 1.3–1.4. This method was applied in [20, 21] to take into account radiative cooling in hydrodynamical binaries. In the explicit approach, the radiative cooling of an optically thin plasma in ionization balance is directly calculated. This technique is described in detail by Cox and Daltabuilt [22]. Such a radiative-cooling model was first applied by Blondin *et al.* [17].

The interaction between the flow and the atmosphere of the accretor represents another difficult problem when calculating the flow patterns in close binaries with normal components. This interaction occurs if the radius of the accretor exceeds 0.1 (here and below, all dimensions are given in units of the component separation). We described this interaction using the coefficient f_v, which decreases the gas velocity at the surface of this star compared to adjacent areas. For example, if the gas velocity in the vicinity of the accretor is V, it will be $f_v \times V$ at the stellar surface. If this factor is zero, there is no accretion onto the star (the boundary conditions correspond to total reflection of the flow); if f_v is unity, the boundary conditions correspond to the total absorption of the flow (free motion of matter through the boundary of the accretor). Intermediate values imply partial absorption of the flow. In our calculations, we set $f_v = 0.5$. The initial model was specified as is described by Nazarenko *et al.* [15]. The initial structure of the flow inside the Roche lobe of the donor in the vicinity of the inner Lagrange point was calculated applying barotropic conditions; i.e., we assumed that the contours of equal density and temperature and the total potential of the binary coincided. Beyond this lobe, the contours of equal density and temperature are assumed to be parallel to the Roche-lobe surface. The transition from the donor atmosphere to the remainder of the computation domain was taken to be continuous.

The boundary conditions in the computation domain were specified as follows. We specified the gradients of physical parameters to be zero at the outer boundary of the domain, enabling the free flow of gas through this boundary. Physical parameters were taken to have their initial values both deep in the Roche lobe of the donor and at the Roche-lobe surface.

The gas flow was calculated over 17–18 orbital periods in order to ensure that a well established steady state was reached in the computation domain.

4. CALCULATIONS TAKING INTO ACCOUNT RADIATIVE COOLING IMPLICITLY

4.1. Model for U Cep

Recall that we modeled the overfilling of the Roche lobe by the donor so as to obtain a mass-transfer rate in the vicinity of L_1 that was close to the observed value, about $10^{-6} M_\odot$/yr. This corresponds to the position of the fifth layer of the atmosphere being at L_1 (the donor atmosphere consists of 10 layers in all), with the degree of overfilling being equal to $\Delta r/r =$

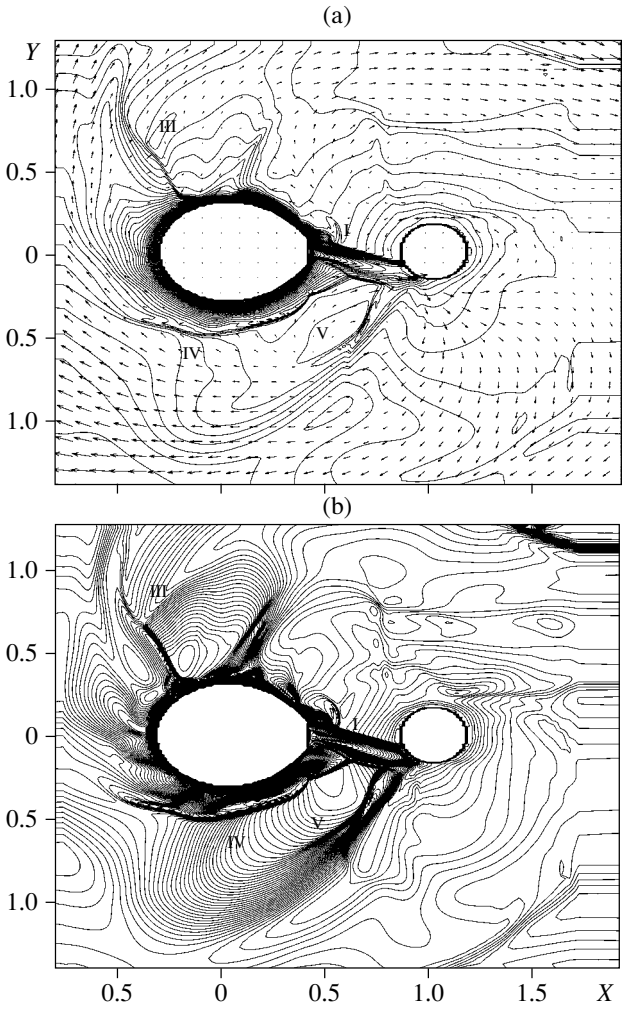

Fig. 1. (a) Contours of equal density and velocity vectors and (b) contours of equal temperature for U Cep with radiative cooling implicitly taken into account.

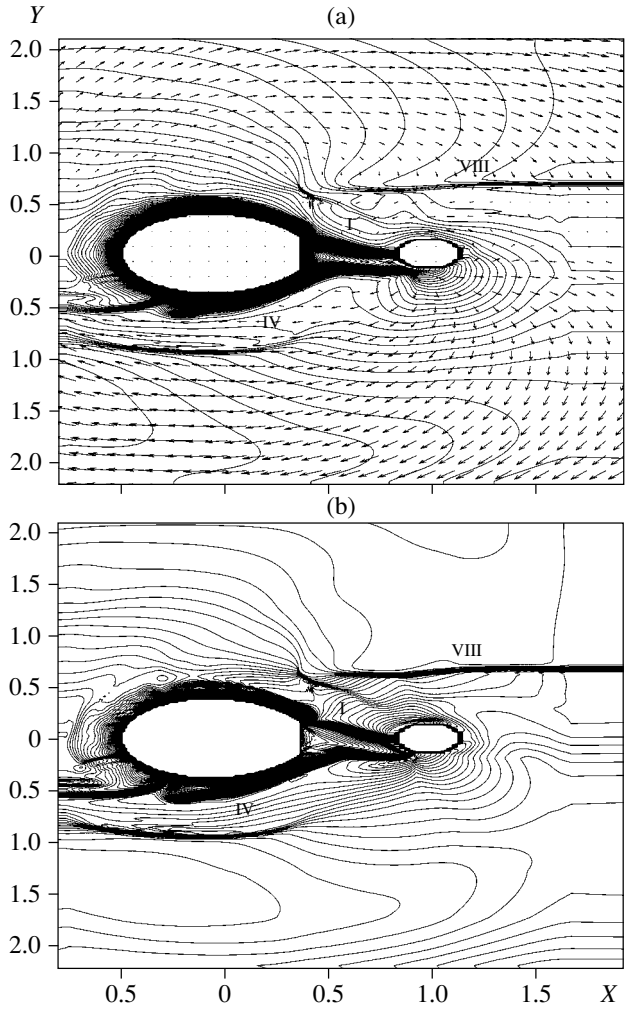

Fig. 2. Same as Fig. 1 for V373 Cas.

0.04 in relative units. In the steady state, the density and temperature at L_1 are equal to 5×10^{15} cm^{-3} and 25 000 K. Figure 1 presents the calculated contours of equal density and temperature and the velocity field. We can see that the flow is not strongly deflected by the Coriolis force either at L_1 or in the course of its motion within the Roche lobe of the primary (the deflection is 0.15), consistent with the model of Lubow and Shu [13]. When $f_v = 0.5$, the flow reflected from the accretor forms two structures. The gas that circulates around the accretor forms the accretion disk, while the gas moving around both components of the binary forms a very dense and extended outer envelope. The number density in the outer envelope varies in the interval 10^{10}–10^{12} cm^{-3}. This comes about due to the almost head-on collision between the flow and the atmosphere of the accreting star. Figure 1 indicates that the morphology of the gas motions in a close binary with normal components differs substantially from those in a system in which the accretor is a compact object [21, 23]. A feature in common for gas motions in systems with normal components and with compact objects is a gas flow from the disk to the point L_3. Further, the gas moves from L_3 to L_1 and around both components of the binary, along the outer boundary of the computation domain, in the direction opposite that of the binary's rotation. The bulk of the gas from the flow moves into the outer envelope, resulting in the formation of a small accretion disk. This contradicts observations, which provide evidence for a quasi-steady-state accretion disk in the system. Since the results of the calculations depend very strongly on the component-mass ratio, which is known fairly well for U Cep, it is possible that, in this case, the observations were interpreted in the standard way, and the outer envelope was taken to be part of the accretion disk.

The gas motions in U Cep give rise to a system of shock waves, which are in many respects similar to

those in binaries with compact objects. Here, we use the same notation for shocks as that of Nazarenko *et al.* [16]. Figure 1 displays shock I, which originates during the collision between the gas moving around the accretor and the gas flow moving from L_1. The spiral shock II is absent from the model for this system. Shock III is located at orbital phases 0.05–0.15 (above the donor) and is formed during the interaction of the gas reflected from the region of shock I and the gas moving along the surface of the donor. Shock IV is formed during the interaction of the gas moving from L_3 and the gas located in the vicinity of the surface of the secondary. Shock V is formed during the interaction of the gas moving from L_3 and the gas reflected from the outer boundary of the flow. Thus, the gas-dynamical model for U Cep predicts the formation of the same shocks as in binaries with compact objects, with the exception of shock II, which is not seen here for a natural reason—it is situated near the surface of the accretor.

The mass transfer in U Cep leads to the formation of high-temperature regions located in the zone of the interaction between the flow and the accretor atmosphere, and also near shock I near the accretor's surface. This region is believed to contribute substantially to the observed Hα emission [24]. Another high-temperature region is part of the accretion disk close to the surface of the accretor and shock V.

4.2. Model for V373 Cas

As was noted above, this system is at the beginning of the first mass-transfer phase, so that its component-mass ratio is smaller than unity. From a hydrodynamical point of view, this implies that the flow moves within the relatively small Roche lobe of the accretor, resulting in an even more direct impact of the flow onto the surface of the star than in the previous case. A mass-transfer rate of the order of $10^{-6} M_\odot$/yr is reached if L_1 lies in the fourth layer of the donor atmosphere and the degree of overfilling in relative units is $\Delta r/r = 0.03$. The temperature and number density at L_1 are 20 000 K and 10^{15} cm^{-3}. In the course of the interaction between the flow and the atmosphere of the accreting star, the matter in the flow is reflected from the stellar surface directly back into the flow. Figure 2 presents the contours of equal density and temperature and the velocity field for V373 Cas. We can see that the gas from the accretor immediately turns toward L_1 along the lower boundary of the flow, giving rise to the very strong shock IV near the surface of the donor. Some of the gas from the flow moves along the surface of the accretor to L_2 and then turns back to L_1 in the direction opposite to the system's rotation. The

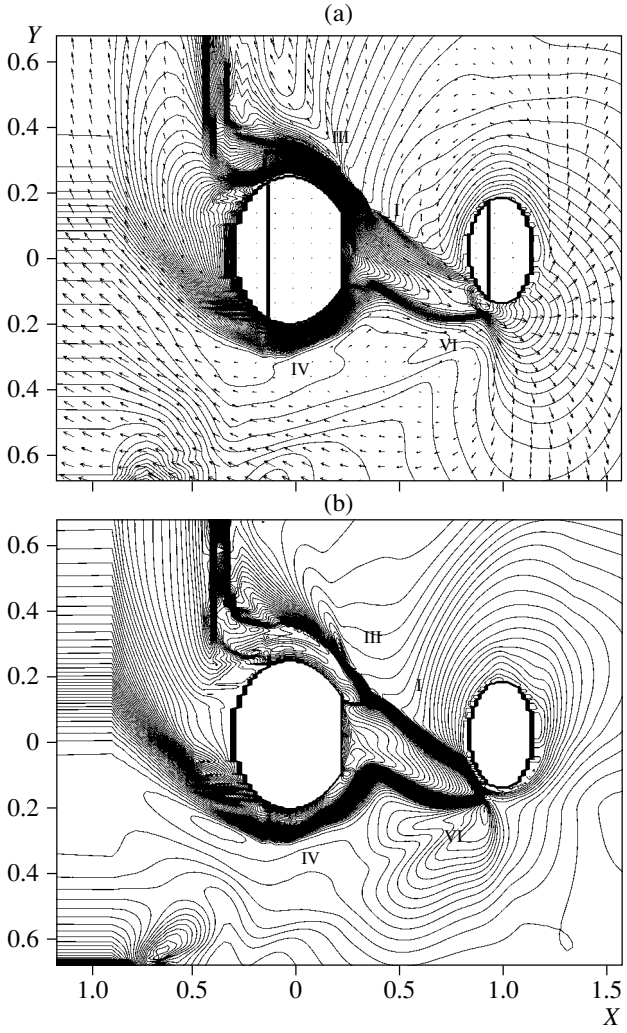

Fig. 3. Same as Fig. 1 for RZ Sct.

remainder of the gas flow circulates around the accretor and forms a small accretion disk with a number density of 10^{12}–10^{14} cm^{-3}. The flow collides with the gas moving around the accretor, resulting in the formation of shock I. The gas at the outer boundary of the computation domain moving opposite to the binary rotation collides with the gas from the accretion disk, forming shock VIII. This last shock is similar to shock II described by Bisikalo *et al.* [21, 23]. We can see from Fig. 2 that shocks III and V are not seen in this binary, probably due to the more direct impact of the flow onto the atmosphere of the accretor compared to the previous case.

Thus, the morphology of the gas motions in V373 Cas is as follows: the system contains a small, dense accretion disk and extended common envelope surrounding the disk and both components of the binary. High-temperature regions in the model for V373 Cas are located in part of the accretion disk very close to the accretor surface, and also near that

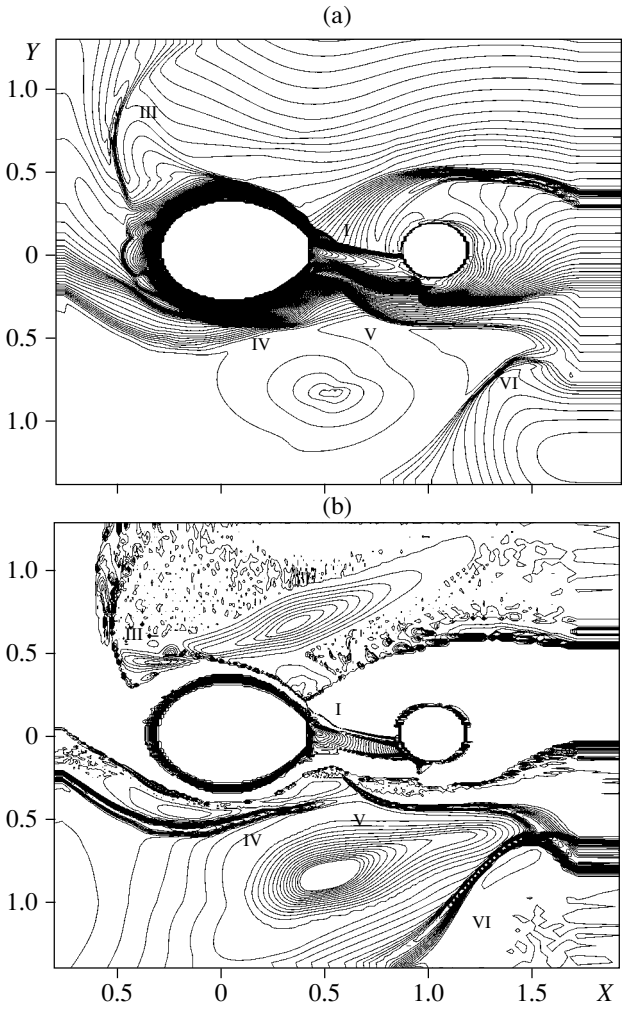

Fig. 4. Same as Fig. 1 with radiative cooling explicitly taken into account.

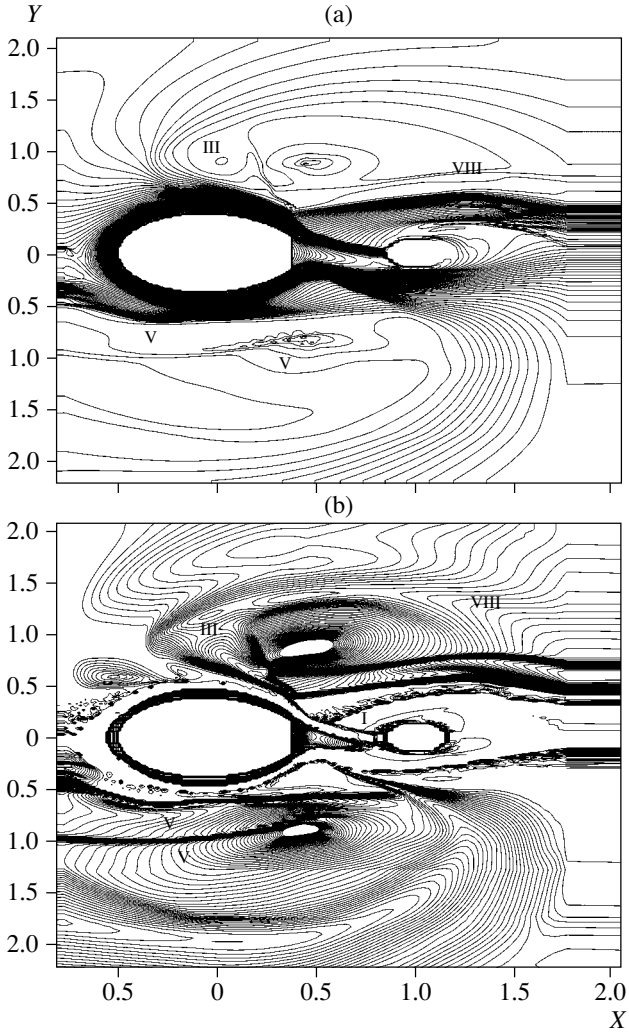

Fig. 5. Same as Fig. 4 for V373 Cas.

parts of shock I that are close to the surface of the accretor, in the region of shock IV below the donor, and in the region of shock VIII.

4.3. Model for RZ Sct

The degree of overfilling of this binary in relative units is $\Delta r/r = 0.035$. In the steady state, the temperature and number density at L_1 are 56 000 K and 10^{15} cm^{-3}. This yields a mass-transfer rate through L_1 of the order of $10^{-6} M_\odot$/yr. Figure 3 presents the resulting contours of constant density and temperature and the velocity field. We can see that the gas morphology in this system is similar to that for U Cep, although the interaction between the flow and the atmosphere of the accretor is substantially different. In RZ Sct, there is a grazing impact of the flow in the disk of the star, due to the high component-mass ratio. The deflection of the flow from the line joining the component centers under the action of the Coriolis force is 0.23 (it is equal to 0.22 in the model of Lubow and Shu [13] for this mass ratio). In this case, only a small fraction of the gas from the flow is reflected from the atmosphere of the accretor, resulting in the formation of a large accretion disk and a less dense outer envelope. As we can see from Fig. 3a, the flow is attracted by the accretor, leading to an increase in the area of the interaction between the flow and the stellar atmosphere. As a result, an accretion disk with a very inhomogeneous density distribution is formed. The maximum density is reached at phases 0.65–0.7, while the accretion disk is very extended at phases 0.25–0.35. The main difference from the behavior seen for U Cep is the presence of a flow from L_2 (an L_2 flow). In this binary, matter flows out from the donor simultaneously through both L_1 and L_2, presumably due to the high component-mass ratio. As in U Cep, there is gas in RZ Sct that moves from L_3 to L_1 and subsequently along the outer boundary of the computation domain, where it rotates opposite to

the rotation of the system. High-temperature regions are located in the zone of interaction between the flow and stellar atmosphere, as well as in part of the accretion disk near its surface. All the matter flowing from L_1 is captured in the Roche lobe of the accretor (conservative mass transfer), whereas all the matter flowing from L_2 leaves the system. The amount of matter flowing from L_2 is about 40% of the amount flowing from L_1. Thus, the mass transfer in RZ Sct is non-conservative. In the course of its subsequent movement, the flow from L_2 is deflected by Coriolis forces and expands, producing a wind in a zone close to the orbital plane, with a velocity of 100–300 km/s in the radial direction. Such winds are characteristic only of close binaries, and we have called them binary-system winds. The shocks in RZ Sct resemble those in the other systems (I, III, IV); however, they differ in the presence of shock VI, which is associated with the shear interaction between the flow and the stellar atmosphere.

5. EXPLICIT ACCOUNT OF RADIATIVE COOLING

The main result of our calculations with an explicit treatment of radiative cooling is a very low temperature of the flow. The temperature in the computation domain decreases to 1000 K (the lower limit of the temperature in the radiative-cooling model used) everywhere except in the region of interaction between the flow and accretor and part of the accretion disk close to the surface of the accretor. Such substantial cooling is due to the high gas density in the outer envelope and disk (of the order of 10^{10}–10^{13} cm^{-3}), which results in the radiation of nearly all the energy from the envelope. Figures 4–6 present contours of constant density and temperature for the studied systems when radiative cooling is taken into account explicitly.

The morphology of the gas and interaction between the flow and the star are the same as in an implicit treatment of cooling; i.e., the relationship between the densities and sizes of the accretion disk and the common envelope is the same. However, the shapes of the accretion disk and outer envelope are substantially different. The mass transfer in the case of an explicit account of radiative cooling is characterized by a clumpy structure of the accretion disk and outer envelope, probably associated with the Rayleigh–Taylor instability. A characteristic feature of the formation of the envelopes with explicit account of radiative cooling is the role of the Lagrange points L_4 and L_5 (in V373 Cas and U Cep). The system of shock waves that form and their origin are similar to those for the case of implicit radiative cooling.

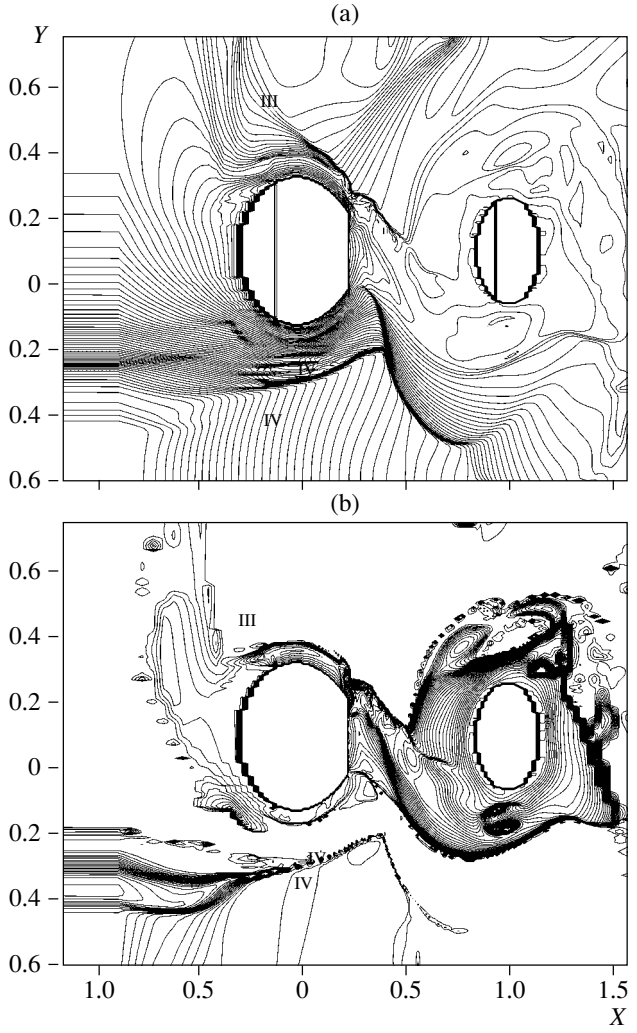

Fig. 6. Same as Fig. 4 for RZ Sct.

To represent the matter distribution in the studied systems, we determined the number of particles along the line of sight as a function of the orbital phase in steps of 5°. The results are presented in Fig. 7, which clearly shows the asymmetry of the envelopes in U Cep and RZ Sct.

6. CONSERVATIVENESS OF THE MASS TRANSFER

In the studied systems, the mass-transfer rate was assumed to be of the order of $10^{-6} M_\odot$/yr, close to the observed value. The mass-transfer rates are the same for both treatments of radiative cooling. The mass transfer in V373 Cas and U Cep is nearly conservative: the ratios of the amount of matter flowing from the system and that flowing from L_1 are $\beta = 0.01$ and 0.03, respectively. Due to the high component-mass ratio in RZ Sct, L_2 is located close to the surface of the donor, resulting in an appreciable outflow of

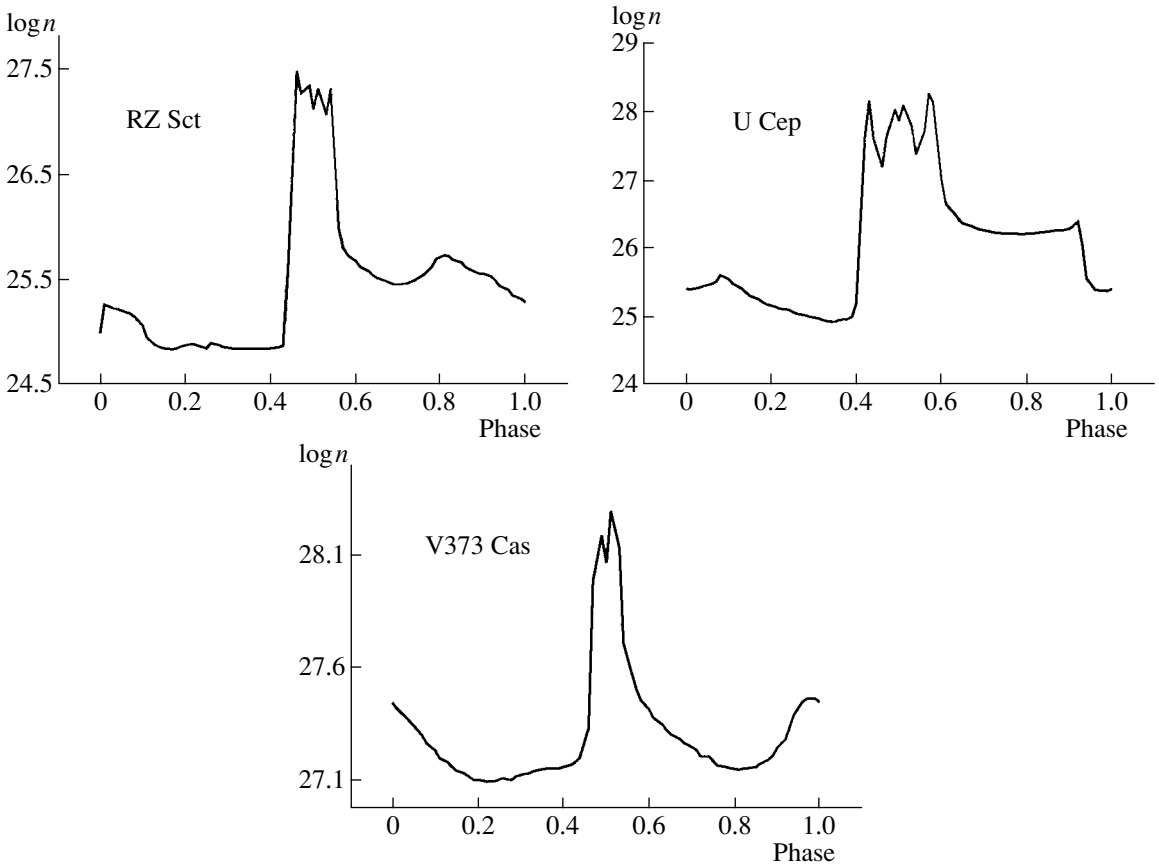

Fig. 7. Orbital phase distribution of the number of particles (on a logarithmic scale) along the line of sight for the studied systems.

matter from this point, which is subsequently lost from the system. In RZ Sct, $\beta = 0.4$, indicating that the mass transfer is non-conservative. It is important for the theory of binary evolution to determine the parameter $\psi = \partial(\ln J)/\partial(\ln M)$, where J is the total angular momentum of the system and $M = M_1 + M_2$, since, along with β, this parameter appears in the equation determining the amount of angular momentum lost by the system. Table 2 presents values of ψ for the three systems. Given β and ψ, we can determine the specific angular momentum β_J lost by the system in the course of the mass transfer. This parameter is also presented in Table 2 (in units of the angular momentum of the system).

7. CALCULATION OF THE ASYNCHRONICITY OF THE ROTATION OF THE BINARY COMPONENTS DUE TO ACCRETION AND OUTFLOW FROM L_1

Adopting the thickness of the disk in the z direction to be equal to the radius of the accreting star, we can calculate the angular momentum gained by the accretor over some time (for example, a year) in terms of the momentum flux through the surface of the star. In this case, we represented the accretor's surface as a cylinder with a height equal to the thickness of the disk and a radius equal to that of the star. The time for variations in the angular velocity of the accretor can be calculated in two different ways:

(1) assuming that the added angular momentum is redistributed over the entire star instantaneously;

(2) assuming that the added angular momentum is initially redistributed only over the outer layers, then slowly (on the evolutionary timescale) over the entire star.

The calculations of the variations of the accretor's angular velocity were very similar in the two cases, except that, in the second case, we used the angular momentum of the upper layers of the star rather than the total angular momentum to determine the variation timescale. If the flux of angular momentum described above is known, we can obtain a new angular momentum for the accretor, and thus calculate the change in the angular velocity of this star very easily, assuming a constant rotational velocity inside the star. We believe this approximation to be reasonable since the observational values of $\dot{M}$, and accordingly

$\dot{\omega}$, are known only to order of magnitude (more accurate determinations are impossible given the current state of both theory and observations). Thus, adding the angular-momentum flux to the previous angular momentum of the accretor (taking into account the added mass and the mass–radius relation for binaries, $M \propto R^{3.4}$, the radius varies as $1 + (M_{\rm add}/M_{\rm acc})^{-3.4}$, where $M_{\rm add}$ is the mass falling onto the star per year), we will obtain the change of the accretor's angular velocity of rotation per year and, accordingly, the rotational asynchronicity acquired by the accretor. When dividing the new angular momentum into the new moment of inertia of the accretor, we can estimate the accretor's moment of inertia using the radial density distribution for a star of the given mass taken from theoretical calculations of the internal composition of stars [25]. In our approach to the redistribution of the added angular momentum, we do not consider this process in detail.

Variations of the rotational angular velocity of the donor were calculated in the same way, except that we calculated the new angular momentum of the donor using the angular-momentum flux through the vicinity of L_1. The fourth column in Table 2 contains the timescale for variations of the accretor's angular velocity for both cases of angular-momentum redistribution. Table 2 also presents the synchronization timescales for both the accretor and donor calculated using the formula of Claret and Gimenez [2] (Column 2), the change in the donor's angular velocity (column 3), $\Delta P/P$ (Column 5), ψ (Column 6), the specific angular momentum of the matter lost by the system β_J in units of the system's angular momentum (Column 7), and the mass-transfer rate (Column 8). The timescales are given for variations of the angular velocities of the donor and accretor by a factor of two.

The $\Delta P/P$ values were calculated using the formula of Tsesevich [26]. Table 2 indicates that, in all three cases, the timescale for variations of the angular velocity of the accretor is smaller than the synchronization timescale only when the redistribution of the accreted angular momentum over the star is slow. Table 2 also provides evidence that the asynchronous rotation of the donor in V373 Cas cannot be due to the flow of matter through L_1.

8. DISCUSSION

We have carried out two-dimensional modeling of the mass transfer in three semidetached binaries in various stages of their evolution. These systems display asynchronous rotation of their components and high mass-transfer rates. Since the systems are semidetached, the theory of tidal interactions in close binaries predicts that the rotation of their components and their revolution should become synchronized. One possible explanation for the asynchronicity that has been put forth by some authors (see Introduction) is a high rate of mass loss through L_1 by the donor and the subsequent accretion of this matter onto the accretor. Our calculations have shown that asynchronous rotation of the accretor can be explained in this way only if the added angular momentum is redistributed over the star over a time interval exceeding the synchronization time. This can occur for specific viscosities lower than 10^{10} cm^2/s. It is very difficult to estimate the viscosities of stars, or at least upper limits for them. If we suppose that the viscosity is related, for example, to turbulence, the specific viscosity will be of the order of $V_s \times \Delta r_s$, where V_s is the scale of the velocity and Δr_s is the spatial scale of the turbulent inhomogeneity. Assuming $V_s \sim$ 10 km/s and $\Delta r_s \sim 0.001$ for the stellar radius, we will obtain the specific viscosity 10^{14} cm^2/s. In any case, it seems unlikely that the stellar viscosity could be lower than 10^{10} cm^2/s. Consequently, we conclude that accretion is not the origin of the asynchronous rotation of the accretor, even if the redistribution of the angular momentum is slow. Asynchronous rotation of a mass-losing donor, likewise, cannot originate due to the matter outflow through the L_1 point. These two conclusions mean that it is impossible to explain the asynchronous rotation of the stars in these binaries as a consequence of accretion or matter outflows. Instead, we can consider the following possible explanations of the observed asynchronicity of the accretor's rotation.

(1) These systems are in a stage of their evolution when synchronization after the first mass-transfer phase is not yet complete; this is possible for RZ Sct and U Cep.

(2) The rotational velocities determined for the accreting stars correspond to the surrounding accretion disk rather than to the star itself. This could be the case for RZ Sct and U Cep. This is a likely explanation since, for example, different rotational velocities are found using different spectral lines in RZ Sct. Our calculations indicate that the density in the disk is close to that in the stellar atmosphere, supporting the possibility that spectral lines are formed in the disk.

(3) The mass-transfer rate in a close binary is estimated primarily from $\Delta P/P$, which is the result of several processes in the system. This estimate yields only a lower limit for this value, especially for systems with non-conservative mass transfer; therefore, it may need to be increased by an order of magnitude. In this case, the timescale for the redistribution of the angular momentum over the star will be shorter than the synchronization timescale, which could then

Table 2. Parameters of mass transfer in the studied systems

System	Synchronization time, 10^5 years	Donor timescale, 10^5 years	Accretor timescale, 10^5 years	$\Delta P/P$	ψ	$\beta_J, 10^{-6}$	$\dot{M}, M_\odot$/yr
U Cep	40.0		45.10	3.0×10^{-6}	0.1	1.6	4.2×10^{-6}
	4.0	3.56	4.37				
V373 Cas	12.0		900.00	-6.2×10^{-6}	0.3	1.5	9.8×10^{-7}
	1.2	87.75	0.85				
RZ Sct	300.0		7900.00	1.5×10^{-6}	4.2	0.8	2.1×10^{-6}
	31.2	258.78	17.48				

explain the asynchronous rotation of the accretor in U Cep.

Thus, in our model for the mass transfer in three semidetached systems, the observed asynchronicity of the components' rotation cannot be unambiguously explained by high mass-transfer rates, as has been suggested in previous studies. It is even more difficult to explain asynchronicity in other systems, whose primaries are rotating asynchronously and which display substantially lower mass-transfer rates. We consider the following two explanations for asynchronicity in Algol-like systems to be the most likely:

(1) the phase of rapid mass transfer finished relatively recently (RZ Sct and U Cep) so that the synchronization is not yet complete,

(2) the accretion disks in these systems are optically thick, so that it is possible for spectral lines to originate in these disks as well as at the stellar surface.

9. CONCLUSION

Our two-dimensional hydrodynamical calculations of mass exchange in three semidetached binaries have yielded the detailed pattern of the gas flows and shock waves in these systems. This indicates that the numerical technique used in our calculations does not smear hydrodynamical details. The technique also makes it possible to obtain results over a relatively short time using personal computers, in contrast to methods used in other similar calculations. For these reasons, we believe we have obtained a reasonably good gas-dynamical model for the mass transfer in binary systems.

Overall, the flow of matter in the Roche lobe of the primary in our calculations is similar to that in the model of Lubow and Shu [13]; however, we have included the influence of the Coriolis force in the vicinity of L_1. This influence is smaller in our model than in the model of Lubow and Shu [13]. The interaction between the flow and the surface of the accreting star depends on the component-mass ratio, the radius of the accretor, and the radius of the flow. In V373 Cas, the component-mass ratio is close to unity, resulting in a direct impact of the flow onto the atmosphere of the accretor. In U Cep, the flow is deflected from the line connecting the component centers, but only slightly. In RZ Sct, the component-mass ratio exceeds unity, leading to a tangential interaction between the flow and the atmosphere of the accretor.

Summarizing the results of our modeling of mass transfer in binaries, we can draw the following conclusions.

1. The morphologies of the gas motions are similar for all three binary systems considered. Consequently, we suggest that other interacting binaries whose components are not compact objects (black holes, neutron stars, or white dwarfs) also display such morphologies. The flow moving from L_1 collides with the surface of the accretor and is reflected from it. Due to this interaction, the flow ceases its motion, and the gas is reflected from the surface of the star and is splashed in various directions, resulting in the formation of an accretion disk (in RZ Sct) and an envelope surrounding both components of the system (in V373 Cas, U Cep). In the case of a direct impact of the flow on the surface of the star (i.e., for component-mass ratios close to unity), the radius of the accretion disk is small, while the common envelope is very dense and extended. In the case of a grazing impact of the flow on the stellar surface (for component-mass ratios substantially exceeding unity), the accretion disk is comparable to the size of the Roche lobe of the accreting star, while the outer envelope is less dense and extended. In the outer envelope, the gas moves from the disk and from the flow toward L_3, then back to L_1, and further rotates around both components in the direction opposite to the system's rotation. In the

course of its interaction with the stellar surface, the flow forms an almost cylindrical accretion disk (reproducing the shape of the stellar surface), in contrast to binaries with compact objects, where the flow moving around the compact star forms an elliptical disk with a very complex structure.

2. Taking into account radiative cooling explicitly yields a system of shock waves that depends on the type of interaction between the flow and the stellar surface:

(a) in the case of a grazing or indirect impact, we obtain shocks I, III, IV, and V (RZ Sct and U Cep);

(b) when the impact of the flow with the stellar surface is direct, the number of shocks is smaller, since the flow is reflected from the surface at an angle of almost 90°, forming shocks I, IV, and VIII.

3. When radiative cooling is taken into account implicitly, three high-temperature regions are formed:

(a) the region of interaction between the flow and surface of the accretor above and below the flow (surrounding the region of interaction between the flow and the star), and part of the accretion disk near the stellar surface, where the gas impacts and is reflected;

(b) the region where the gas moving from the disk and flow toward L_3 is reflected from the secondary and interacts with the flow from L_1 (the region of shocks IV and V);

(c) the region where the gas surrounding both components interacts with the accretion disk (at phases 0.25–0.4).

4. The morphology of the gas in the systems varies only slightly when radiative cooling is taken into account explicitly. In this case, the structure of the gas may be clumpy, apparently due to the Rayleigh–Taylor instability. When radiative cooling is taken into account explicitly, the high-temperature region is the zone of interaction between the flow and the surface of the accretor. Thus, the overall gas motions in a binary depend only slightly on the way in which the radiative cooling is taken into account. Nonetheless, we suggest that an explicit treatment is more correct.

5. The mass transfer in V373 Cas and U Cep is almost completely conservative; i.e., the ratio of the amount of matter leaving the system to the amount flowing from L_1 is 0.01–0.02. In RZ Sct, the mass transfer is non-conservative, and this ratio is equal to 0.4.

REFERENCES

1. I. B. Pustyl'nik, Itogi Nauki Tekh., Ser. Astron. 36 (1989).
2. A. Claret and A. Gimenez, Astron. Astrophys. **296**, 180 (1995).
3. L. V. Glazunova, Pis'ma Astron. Zh. **25**, 545 (1999) [Astron. Lett. **25**, 467 (1999)].
4. L. S. Lyubimkov, T. M. Rachkovskaya, S. I. Rostopchin, and A. E. Tarasov, Astron. Zh. **75**, 355 (1998) [Astron. Rep. **42**, 312 (1998)].
5. M. T. Richards and G. E. Albright, Astrophys. J., Suppl. Ser. **123**, 537 (1999).
6. E. C. Olson and P. B. Etzel, Astron. J. **108**, 262 (1994).
7. E. C. Olson and D. J. Bell, Publ. Astron. Soc. Pac. **101**, 907 (1989).
8. V. G. Karetnikov, Perem. Zvezdy **17**, 479 (1970).
9. J. Mukherjee, G. J. Peters, and R. E. Wilson, Mon. Not. R. Astron. Soc. **283**, 613 (1996).
10. B. J. Burnett and P. B. Etzel, Astron. J. **106**, 1627 (1993).
11. S. O. Selam and O. Demircan, in *Variable Stars as Essential Astrophysical Tools,* Ed. by Cafer Ibanoglu (Kluwer, Boston, 2000), p. 605.
12. R. L. Kurucz, Astrophys. J., Suppl. Ser. **40**, 1 (1979).
13. S. Lubow and d F. H. Shu, Astrophys. J. **198**, 383 (1975).
14. O. M. Belotserkovskiĭ and Yu. M. Davydov, *Large Particles in Gas Dynamics* (Nauka, Moscow, 1982) [in Russian].
15. V. V. Nazarenko, L. V. Glazunova, and V. G. Karetnikov, Astron. Zh. **78**, 525 (2001) [Astron. Rep. **45**, 452 (2001)].
16. V. V. Nazarenko and L. V. Glazunova, Astron. Zh. **80**, 1099 (2003) [Astron. Rep. **47**, 1013 (2003)].
17. J. M. Blondin, M. T. Richards, and M. L. Malinowski, Astrophys. J. **445**, 939 (1995).
18. M. T. Richards and M. A. Ratliff, Astrophys. J. **493**, 326 (1998).
19. D. V. Bisikalo, P. Harmanec, A. A. Boyarchuk, *et al.*, Astron. Astrophys. **353**, 1009 (2000).
20. D. Molteni, G. Belvedere, and G. Lanzafame, Mon. Not. R. Astron. Soc. **249**, 748 (1991).
21. D. V. Bisikalo, A. A. Boyarchuk, V. M. Chechetkin, *et al.,* Astron. Zh. **76**, 985 (1999) [Astron. Rep. **43**, 797 (1999)].
22. D. P. Cox and E. Daltabuilt, Astrophys. J. **167**, 113 (1971).
23. D. V. Bisikalo, A. A. Boyarchuk, V. M. Chechetkin, *et al.*, Mon. Not. R. Astron. Soc. **300**, 39 (1998).
24. M. T. Richards, Astrophys. J. **387**, 329 (1992).
25. Yu. L. Frantsman and V. L. Varshavskiĭ, Nauchn. Inform. Astron. Soveta Akad. Nauk SSSR **6**, 48 (1967).
26. V. P. Tsesevich, *Eclipsing Binaries* (Nauka, Moscow, 1971) [in Russian].

Translated by K. Maslennikov

Astronomy Reports, Vol. 47, No. 12, 2003, pp. 1038–1045. Translated from Astronomicheskiĭ Zhurnal, Vol. 80, No. 12, 2003, pp. 1127–1135.
Original Russian Text Copyright © 2003 by Megn, Braude, Rashkovskiy, Sharykin, Shepelev, Inyutin, Vashchishin, Brazhenko, Bulatsen.

Observational Studies of the Angular Structure of the Radio Galaxy 3C 234 at Decameter Wavelengths

A. V. Megn[1], S. Ya. Braude[1], S. L. Rashkovskiy[1], N. K. Sharykin[1], V. A. Shepelev[1], G. A. Inyutin[1], R. V. Vashchishin[2], A. I. Brazhenko[2], and V. G. Bulatsen[2]

[1] *Institute of Radio Astronomy, Kharkov, Ukraine*

[2] *Gravimetric Observatory of the Institute of Geophysics, Poltava, Ukraine*

Received March 15, 2003; in final form, July 26, 2003

Abstract—An analysis of the angular structure of the radio galaxy 3C 234 at decameter wavelengths based on data obtained on the URAN-1 and URAN-2 interferometers is presented. Four of the five model components that describe the radio-brightness distribution at centimeter wavelengths are observed at decameter wavelengths: two compact components and two neighboring extended components. The fifth, undetected, component is the most extended, and encompasses the central region of the radio source, including the nucleus of the galaxy. Self-absorption is detected in the compact components, whose angular sizes are determined to be $0.27 \pm 0.03''$ (northeast component) and $0.55 \pm 0.05''$ (southwest component), in agreement with direct measurements at centimeter wavelengths. Most of the decameter emission of the radio galaxy is associated with its extended components. © *2003 MAIK "Nauka/Interperiodica"*.

1. INTRODUCTION

The radio source 3C 234 is associated with an optical galaxy with redshift $z = 0.185$ located at the edge of a small group of galaxies. It is a classical FRII steep-spectrum radio source. Its spectral index α at frequencies above 1000 MHz is 0.99 ($S \sim \nu^{-\alpha}$, where S is the spectral flux density and ν the frequency of the radiation). The angular structure of the source is fairly complex, making it impossible to determine the radio-brightness distribution with certainty using measurements at frequencies from 86 to 2695 MHz obtained with radio interferometers capable of determining only the amplitude of the visibility function γ as a function of the baseline D in wavelengths D/λ. It was possible only to establish that the angular dimensions of the radio galaxy are fairly large, of the order of several tens of arcseconds, and that the source consists of no fewer than two components, with their extent being much larger in right ascension than in declination (see, for example, [1]).

More detailed information about the radio structure of 3C 234 was obtained using supersynthesis interferometers. For example, an approximate map of the radio-brightness distribution at 1407 MHz was determined, albeit with relatively low resolution ($23'' \times 47''$ in right ascension and declination, respectively), in one of the earliest studies, carried out on the one-mile Cambridge interferometer [2]. In that map, the source consists of two bright and two less bright compact components surrounded by a region of comparatively weak, extended emission. The bulk of the radiation comes from the outer components to the southwest and northeast, whose centers are separated by $99''$ in position angle 65°. The bright components were estimated to be $15''$–$20''$ in size,[1] while the remaining components were less than $10''$–$20''$ in size. In other similar studies targeted primarily at determining the fine structure of the source radiation [3–7], the sizes and fluxes of the core and the most compact components in the lobes were derived at frequencies from 408 to 15 376 MHz. As a rule, the extended lobes themselves were resolved.

The structure of 3C 234 has been determined most fully via imaging using instruments such as the Very Large Array. Many radio maps of 3C 234 from 1400–15 000 MHz have been made, with resolutions better than $0.3''$ in some cases. For example, Alexander [8] presents maps with resolution $5.3'' \times 8''$ at five frequencies from 1.4 to 14.96 GHz obtained with the VLA and the 5-km Cambridge interferometer. Hardcastle *et al.* [9] present high-dynamic-range images with resolutions of $0.3''$ and $2.3''$ obtained using the VLA at 8.44 GHz; the latter image was obtained using all the configurations of the VLA, which made it possible to correctly represent the extended, low-surface-brightness components with high angular resolution. The hot spots (compact features in

[1] Here and below, the angular sizes of radio components are given in terms of the full width at half maximum (FWHM) for a Gaussian approximation to the brightness distribution.

1063-7729/03/4712-1038$24.00 © 2003 MAIK "Nauka/Interperiodica"

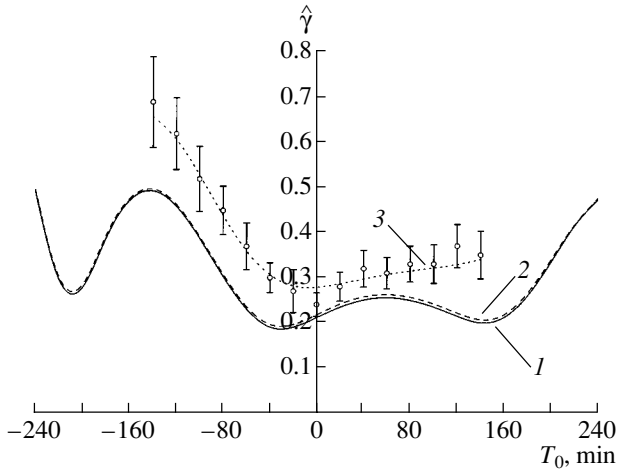

Fig. 1. Amplitudes of the visibility function $\hat{\gamma}$ of 3C 234 at various hour angles T_0 measured by the URAN-1 interferometer at 25 MHz. The circles show the mean-weighted data with the rms deviations $\pm\sigma_{\hat{\gamma}}$. The solid curve *1* shows the calculated dependence $\gamma_p(T_0)$ derived from the 8440 MHz radio map of 3C 234 [9] for an interferometer with the same baseline vector as for the URAN-1 interferometer at 25 MHz. The dashed curve *2* shows the calculated dependence $\gamma_p(T_0)$ for the same interferometer for the five-component model presented in Table 1. The dotted curve *3* shows the calculated dependence $\gamma_p(T_0)$ for the URAN-1 interferometer at 25 MHz for the optimal four-component model fit to the data (Table 2).

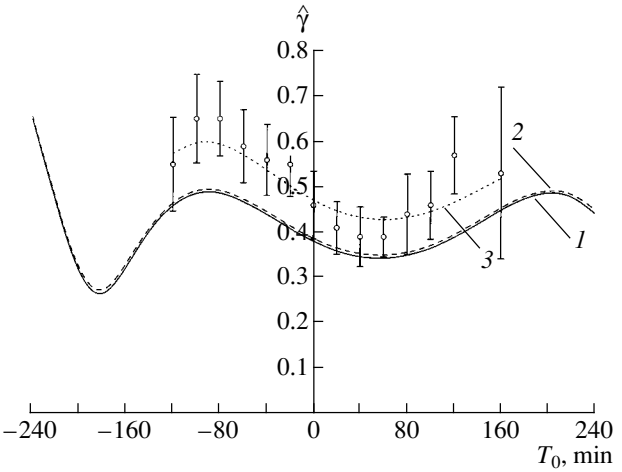

Fig. 2. Same as Fig. 1 for 20 MHz.

the extended lobes) were also studied using the VLA at 15 GHz with resolution $0.21'' \times 0.38''$, and the magnetic fields in these features estimated [10].

Although radio interferometric observations were carried out at 158 MHz [11] and 86 MHz [12], not even the approximate structure of 3C 234 at meter wavelengths has been established. The integrated spectral index decreases to $\alpha = 0.85$ below 1000 MHz, possibly testifying to changes in the angular structure of the object at these frequencies. No interferometric studies had been conducted at decameter wavelengths. However, we expect substantial changes in the radio structure at such wavelengths due to the appreciable influence of cosmic plasma on the radiation and propagation of the electromagnetic waves. It was therefore of considerable interest to derive the angular structure of 3C 234 from radio interferometric observations obtained at the longest wavelengths possible using interferometers located on the surface of the Earth.

2. MEASUREMENTS AND DATA REDUCTION

Our decameter-wavelength observations of 3C 234 were carried out using the URAN-1 and URAN-2 interferometers, with baselines of 42.3 and 152.3 km oriented nearly East–West [13]. The interference fringes were formed by multiplying the signals of the North–South arm of the UTR-2 radio telescope [14] with the signals from the URAN-1 and URAN-2 antennas. The UTR-2 antenna received only a single linearly polarized signal, while the URAN antennas received simultaneously two orthogonal (when receiving from the zenith) linear polarizations. This enabled us to correct for the Faraday effect—rotation of the plane of polarization of the signals in the ionosphere of the Earth and the cosmic plasma through which the signals propagated—which is very substantial at decameter wavelengths. The measurements were conducted simultaneously at 25 and 20 MHz during nighttime in January and February 1996 and March 1999. The measurements yielded the amplitude of the visibility function γ in an interval of about ± 2 hrs on either side of the culmination of the source. Measurements were obtained at various hour angles T_0 in steps of 20 minutes. When processing the recordings, we determined the mean visibility-function amplitudes for each 20-minute interval. The data obtained for each hour angle were averaged over all days of the observations, and we determined the mean-weighted observational values $\hat{\gamma}_{\rm obs}(T_0)$ and their errors $\sigma_{\hat{\gamma}}(T_0)$.

We used the method described in [15] to calibrate the interferometer data, using special digital noise generators to imitate the response of the interferometer to a point source. The calibration generators were synchronized at all points in the interferometers using GPS receivers.

The observational data for 3C 234 are presented in Figs. 1–4. The circles show the mean-weighted amplitudes of the visibility function $\hat{\gamma}_{\rm obs}$ as functions of the hour angle at 25 MHz (Figs. 1, 3) and 20 MHz (Figs. 2, 4) for the URAN-1 (Figs. 1, 2) and

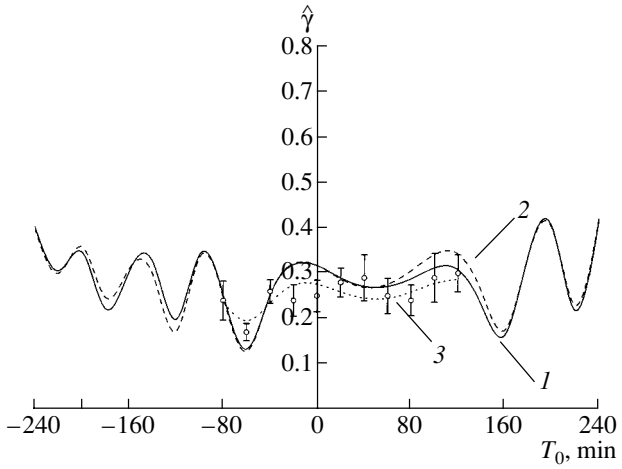

Fig. 3. Same as Fig. 1 for the URAN-2 interferometer.

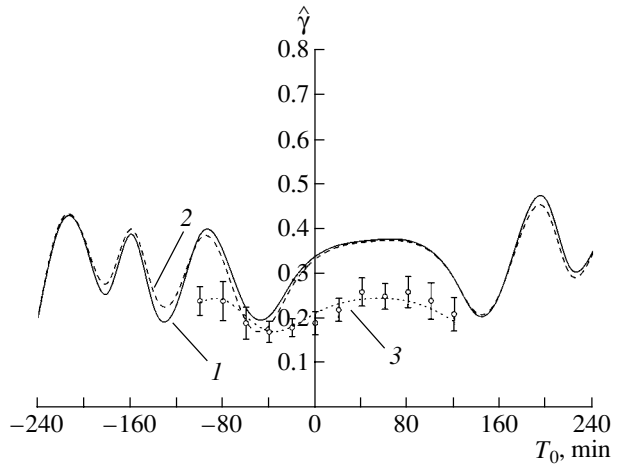

Fig. 4. Same as Fig. 3 for 20 MHz.

URAN-2 (Figs. 3, 4) interferometers. The vertical bars indicate the rms deviations taking into account both the random scatter of the data points and systematic errors. In all cases, we can see oscillations in the observational dependences $\hat{\gamma}_{\rm obs}(T_0)$ demonstrating that the source likewise contains no fewer than two widely separated compact features at decameter wavelengths.

Because the URAN-1 and URAN-2 interferometers did not provide very full coverage of the UV plane, and also due to the absence of phase measurements, we were not able to use standard imaging methods based on calculating the two-dimensional Fourier transforms of the complex visibility functions and iteratively refining the resulting radio maps. Therefore, the most suitable method for deriving the angular structure of the source was by fitting models of the radio brightness distribution to the data and identifying the model whose spatial-frequency spectrum was in best agreement with the measured values $\hat{\gamma}_{\rm obs}$. We used sets of elliptical Gaussians with arbitrary orientations for their position angles as the model components. The model parameters were varied in order to search for good agreement between the calculated and observed dependences of the visibility-function amplitude on the hour angle $\gamma(T_0)$. We adopted the standard χ^2 criterion as a measure of the agreement between the calculated model and observations [16]:

$$\chi^2 = \sum_{i=1}^{N} \left[\frac{\hat{\gamma}_{\rm obs,i}\,(T_0) - \gamma_{{\rm calc},i}\,(T_0)}{\sigma_{\hat{\gamma}_{\rm obs,i}}\,(T_0)} \right]^2, \qquad (1)$$

where N is the number of independent measurements of the mean-weighted visibility-function amplitude $\hat{\gamma}_{\rm obs}$ at different hour angles T_0. The best model for the radio-brightness distribution was taken to be that corresponding to the minimum χ^2, provided there was a sufficiently high probability that it exceeded $W_p(\chi^2 > \chi^2_{\rm min})$, where $p = N - 1 - p_1$ denotes the number of degrees of freedom of the model and p_1 is the number of model parameters fit.

The method used to calculate $\gamma_{\rm calc}(T_0)$ and carry out the model fitting on a personal computer are presented in [17].

It is well known that the derivation of models using only the visibility-function amplitudes can lead to ambiguities in the solutions and requires a large amount of computing time, especially if the structure of the object is complex and the UV coverage is not very complete. In addition, comparisons of the resulting models with higher-frequency maps of the radio source in order to determine variations in the structure with frequency can be only qualitative. We used the following algorithm to simplify the model-fitting procedure and the derivation of quantitative properties of frequency-dependent variations in the source structure. We selected the simplest model that could qualitatively describe the real radio-brightness distribution observed in a higher-frequency map convolved with the resolution of the URAN interferometers. We then used the resulting model as an initial model when fitting our data to search for a suitable model of the radio-brightness distribution at the lower frequencies.

Of course, it is desirable to choose for this purpose a map obtained at a long wavelength that is as close as possible to our own, with resolution better than that of the URAN-2 interferometer and with good reproduction of the low spatial frequencies to which the URAN-1 interferometer is sensitive. A qualitative map of the source with a resolution of 2.3″ adequate for our purpose and good coverage of the UV plane has been obtained only at 8.44 GHz [9]. However, a comparison of maps with lower resolution [8] (obtained in FITS format over the Internet) and the

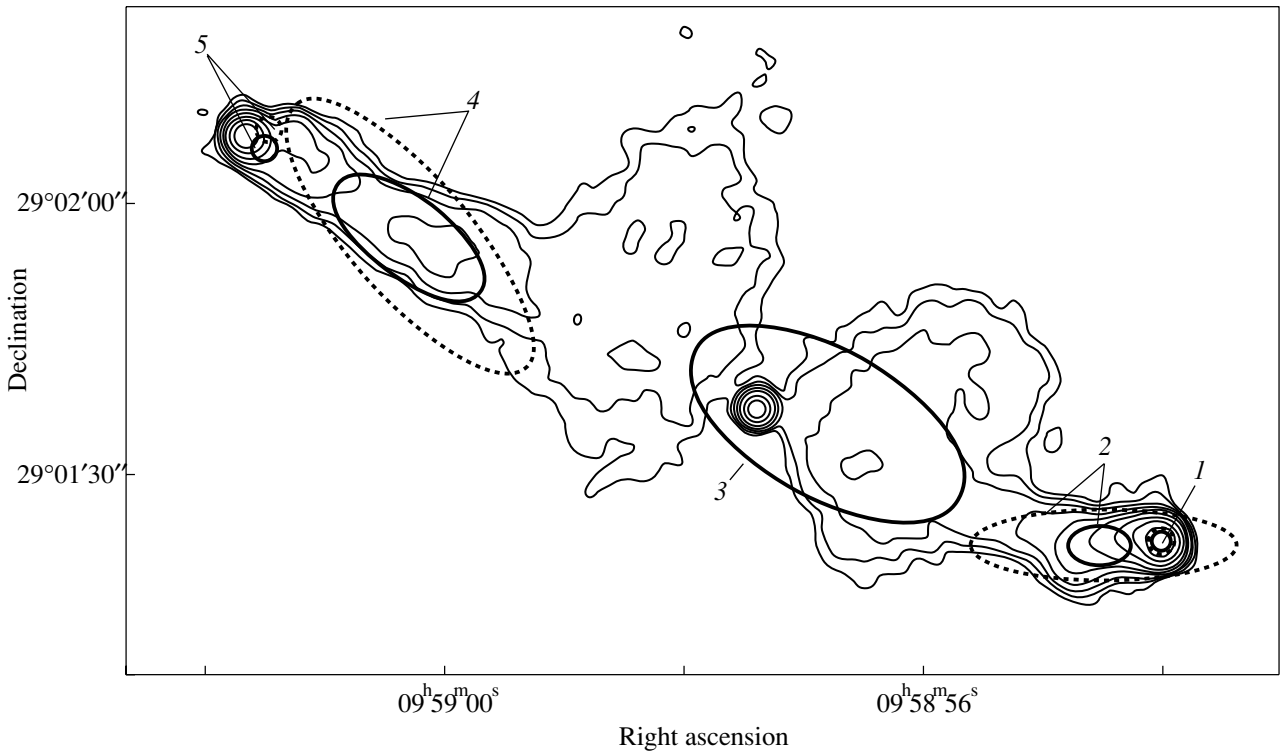

Fig. 5. Best-fit model for the radio-brightness distribution of 3C 234 at 8440 MHz (solid curves) and 25 MHz (dashed curves) superposed on the map of [9] (map included with the agreement of the authors).

map of [9] smoothed to the same lower resolution showed that the source structure remains virtually the same from 1.4 GHz to at least 8.44 GHz (within the resolution of the URAN interferometers). The most substantial variation is a relative decrease in the flux density of the central compact component (core) of the radio galaxy with decreasing frequency; however, we can neglect this component due to the small value of its flux density. Therefore, we used the map of [9] with a resolution of 2.3″, kindly provided by M. Hardcastle, for our subsequent calculations.

3. FITTING OF MODELS OF THE BRIGHTNESS DISTRIBUTION

The solid curves in Figs. 1–4 show the calculated hour-angle dependences of the visibility-function amplitudes that would be observed by the URAN-1 and URAN-2 interferometers if the radio-brightness distribution of 3C 234 was the same at decameter wavelengths as in the map of [9]. We can see in these figures that, although the observational data are in qualitative agreement with the calculated dependences, there are clear quantitative differences between them. For example, the URAN-2 data, which are primarily sensitive to compact features in the brightness distribution, testify to changes in the relative fluxes of the compact components (the depth of the modulation in the hour-angle dependence is smaller at 25 MHz; Fig. 3). The 20-MHz data (Fig. 4) indicate an even more appreciable decrease in the interferometric responses to these components, even with respect to the frequency of 25 MHz possibly associated with an increase in the sizes of the components, or alternatively with a deviation of their spectra from power-law relations. We determined the changes in the brightness distribution at the lower frequencies compared to the higher-frequency maps using the following procedure.

(1) We identified regions in the map of [9] that could be represented as individual Gaussian components. Components 1 and 5 both have sizes of 2.3″, and correspond to the brightness centers of the southwest and northeast lobes of the radio galaxy (these components are actually much smaller, but were taken to be equal to the size of the synthesized beam of the map). Components 2 and 4 represent the parts of the southwest and northeast lobes adjacent to the hot spots. Component 3 represents the extended, diffuse emission around the core and the core itself. The calculations demonstrated that this part of the source was appreciably resolved even by the URAN-1 interferometer, and can be treated like a single component. As was indicated above, the core itself makes only a very small contribution to the total flux of the source.

(2) We then fit the parameters of the five-component for the brightness distribution to obtain the best agreement between the hour-angle dependences $\gamma_{calc}(T_0)$ and the analogous dependences calculated above for the 8.44 GHz map.

The components of the fitted model with their FWHM sizes are shown by the solid ellipses in Fig. 5, and the corresponding parameter values are given in Table 1. We can see that these components are in reasonable agreement in both shape and position with the parts of the map they are meant to describe. Note that the flux densities of the best-fit model components coincide with the flux densities from the corresponding regions in the map to within several percent, and the model dependences $\gamma_{calc}(T_0)$, shown by the dashed curves in Figs. 1–4, agree well with the analogous dependences derived for the map.

We derived models for the radio-brightness distribution of 3C 234 at decameter wavelengths that best described the observational data in Figs. 1–4 using the higher-frequency model as a first approximation. Our calculations showed that even this method for selecting the model does not provide an unambiguous determination of its parameters due to the complexity of the model compared to the incomplete UV coverage of our observations. For example, the uncertainties in the positions and sizes of the components in declination are very large, since the baselines of both interferometers are oriented East–West and the hour-angle range (range of baseline variations) covered is small. The coordinates are determined much more accurately in right ascension. The URAN-2 interferometer substantially resolves all the components except for 1 and 5, enabling us to determine the parameters of these components with good accuracy, especially the separation between them in right ascension and their flux ratio. We were not able to simultaneously derive the parameters of components 2–4, which were only partially resolved by the URAN-1 interferometer. We used therefore their positions obtained for the high-frequency model together with additional considerations about the character of the spectra of these components from 20 to 8440 MHz. The parameters of the resulting models are presented in Table 2, using the same scheme to number the components. The corresponding hour-angle dependences $\gamma_{calc}(T_0)$ are shown by the dotted curves in Figs. 1–4. Note that the model parameters were fit separately for 20 and 25 MHz. The model components are shown by the dashed ellipses in Fig. 5.

In Tables 1 and 2, $\Delta\alpha_i$ and $\Delta\delta_i$ are the coordinates of the component centers in arcseconds relative to the first (southwestern) compact component, S_i/S_0 are the relative spectral flux densities of the components ($S_0 = \sum S_i$ is the total spectral flux density of the radio galaxy), $\Delta\theta_{i\max}$ and $\Delta\theta_{i\min}$ are the maximum and minimum angular sizes of the components in arcseconds, and ψ are the position angles of the major axes of the components in degrees measured from West.

The quantities χ^2_{min} characterizing the agreement between the calculated dependences $\gamma_{calc}(T_0)$ for the models in Table 2 and the observational data are 6.9 for the 20 MHz data and 8.4 for the 25 MHz data, with the probabilities of exceeding them $W_p(\chi^2 > \chi^2_{min})$ being 0.8 and 0.6, respectively.

The errors in the parameters for the low-frequency models in Table 2 were calculated using the method described in [17], and take into account both the scatter in the observational data and the accuracy with which the parameters can be determined given the UV plane coverage. For example, the relative insensitivity of the East–West baselines of the URAN interferometers to the sizes and separations of the components in declination is reflected by the relatively large errors for these parameters presented in the table. The errors in Table 1 for the model fitted to the map reflect only the sensitivity of the model to variations of a given parameter for a specified variation in the projected interferometer baseline.

The sizes of the compact components 1 and 5 measured at 8.4 GHz [9] and 15 GHz [10] are smaller than the resolution of the URAN-2 interferometer, even taking into account their broadening due to scattering of the radio waves on inhomogeneities in the interstellar and interplanetary plasma. Therefore, they were not determined when fitting the model, and Table 2 gives an upper limit for these sizes corresponding to the synthesized beam of the map of [9].

4. ANALYSIS OF THE RESULTING DATA

The models for the radio-brightness distribution we have obtained for the data at centimeter and decameter wavelengths can be used to carry out both qualitative and quantitative comparisons and determine variations of the structure of the radio galaxy with frequency.

First, the coordinates of the centers of the compact components at 8440 MHz and at 25 and 20 MHz virtually coincide. The nominal differences do not exceed the errors, and are probably associated with errors in the measurements and error introduced by the data reduction, as well as deviations of the real angular distributions of the radio emission of the compact components from symmetrical Gaussians. In Fig. 5, the position of component 1 coincides with a compact feature in the southwest lobe of the radio galaxy. However, component 5 does not coincide with the corresponding feature in the northeast lobe, in either the low-frequency or the high-frequency models. This

Table 1. Model radio-brightness distribution for 3C 234 at 8440 MHz

Component	$\Delta\alpha_i$	$\Delta\delta_i$	S_i/S_0	$\Delta\theta_{i\max}$	$\Delta\theta_{i\min}$	ψ
1	$0''$	$0''$	0.29 ± 0.03	$<2.3''$	$<2.3''$	–
2	6.1 ± 0.8	-0.4 ± 1.0	0.14 ± 0.02	6.6 ± 3.4	3.9 ± 1.4	$0 \pm 30°$
3	37.3 ± 2.6	14.7 ± 5.4	0.31 ± 0.02	32.5 ± 6.9	15.6 ± 4.2	28.9 ± 36
4	83.7 ± 2.9	34.2 ± 4.6	0.15 ± 0.03	19.1 ± 7.8	6.5 ± 3.9	38.7 ± 30
5	99.4 ± 1.0	43.6 ± 1.1	0.11 ± 0.01	<2.3	<2.3	–

Table 2. Model radio-brightness distribution for 3C 234 at decameter wavelengths

ν, MHz	Component	$\Delta\alpha_i$	$\Delta\delta_i$	S_i/S_0	$\Delta\theta_{i\max}$	$\Delta\theta_{i\min}$	ψ
20	1	$0''$	$0''$	0.22 ± 0.03	$<2.3''$	$<2.3''$	–
	2	6.1 ± 4.0	-0.4 ± 13	0.33 ± 0.06	35.7 ± 6.0	9.2 ± 2.6	$0 \pm 40°$
	4	83.7 ± 6.0	34 ± 20	0.4 ± 0.05	37.0 ± 10	10.9 ± 4.5	51.8 ± 25
	5	96.3 ± 2.2	46 ± 14	0.05 ± 0.02	<2.3	<2.3	–
25	1	0	0	0.25 ± 0.03	<2.3	<2.3	–
	2	6.1 ± 4.6	-0.4 ± 12	0.33 ± 0.06	29.6 ± 6.5	7.6 ± 2.8	$0 \pm 36°$
	4	83.7 ± 3.5	34 ± 13.5	0.36 ± 0.05	37 ± 6.0	10.9 ± 2.6	51.8 ± 27.5
	5	98.8 ± 1.3	45.6 ± 7.7	0.06 ± 0.02	<2.3	<2.3	–

is due to the complexity of the brightness distribution in the regions of these features, which has not been taken into account in our models.

Second, the fraction of the total radio flux density contributed by the compact components 1 and 5 falls off with decreasing frequency and, as will be shown below, their low-frequency flux densities are lower than predicted by an extrapolation from higher frequencies assuming a power law with a constant spectral index.

Third, the sizes of components 2 and 4 have grown at decameter wavelengths by factors of two to four compared to their sizes at centimeter wavelengths. Their contribution to the total flux density grows from 30% at centimeter wavelengths to nearly 70% at decameter wavelengths, obviously indicating an increase in the spectral indices of these components with decreasing frequency.

Fourth, our most surprising result is the absence of component 3, which represents the diffuse, extended regions of emission near the core of the radio galaxy, in the models for the radio-brightness distribution of 3C 234 at decameter wavelengths.

Let us compare the derived decameter-wavelength flux densities of the components with their well known spectral characteristics at frequencies above 1000 MHz. We can use the maps of [8, 9] to derive the flux densities in regions of the source corresponding to the model components and their spectra from 1.4 to 8.44 GHz. The ratios of the fluxes of the various components remain virtually constant in this frequency range; i.e., the spectral indices of the components are all approximately the same, and are approximately equal to the integrated spectral index of the source to 1000 MHz. The only exception is the core of the radio galaxy. Using these data, our own measurements, and other measurements taken from the literature (those of [18] for the integrated flux densities and those of [3–7, 19] for the flux densities of individual components), we have constructed the integrated spectrum of the source from 14.6 to 15 375 MHz (solid curves in Fig. 6), as well as the radio spectra of the southwestern (curve *1*) and northeastern (curve *5*) components from 408 to 15 375 MHz. The numbers labeling the spectra in Fig. 6 correspond to the component labels in Tables 1 and 2. At frequencies below 408 MHz, where no data are available, proposed continuations of the spectra are shown by dashed curves. The asterisks and crosses denote our own data presented in this paper.

We can see in Fig. 6 that our flux densities for components 1 and 5 lie substantially below the dependences extrapolated from the higher frequencies assuming a constant spectral index. If we suppose

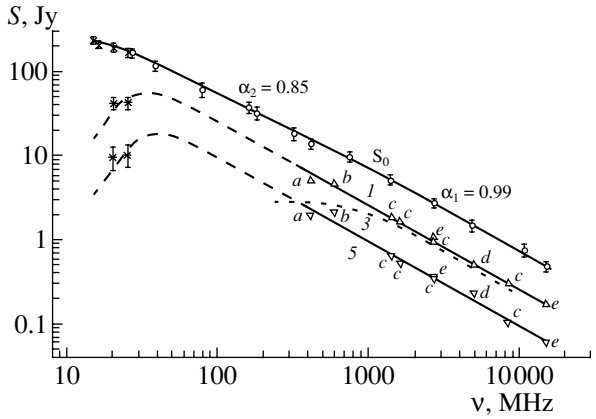

Fig. 6. Spectral characteristics of the radio galaxy 3C 234. S_0 shows the frequency dependence of the total spectral flux density of the radio galaxy, $S_0(\nu)$ (the circles show the data of [18] and the crosses the data of this paper). Curves *1*, *3*, and *5* show the frequency dependences of the spectral flux densities of the southwestern compact component (upward triangles), the northeastern compact component (downward triangles), and the central extended component (dotted curve); the letters denote the corresponding references (a = [3], b = [7], c = [8, 9], d = [4], e = [5, 6]). The asterisks show our measurements of the spectral flux densities of the southwestern and northeastern components.

that self-absorption is occurring in these compact components, we can recalculate the spectra taking this physical process into account.

In the presence of self-absorption, the frequency of the maximum radiation ν_m and the maximum spectral flux density S_m are given by the relations [20]

$$\nu_m = \left[\frac{S_1 \nu_1^{\alpha} H_{\perp}^{1/2} (1+z)^{1/2}}{\tau_0 \Delta\theta^2}\right]^{2/2\alpha+5}, \qquad (2)$$

$$S_m = S_1 \left(\frac{\nu_1}{\nu_m}\right)^{\alpha} \frac{1-e^{-\tau_0}}{\tau_0}, \qquad (3)$$

where S_1 is the spectral flux density of the component (in Jy) at frequency ν_1 (in MHz), $H_{\perp}$ is the component of the magnetic field in the compact component perpendicular to the direction of motion of the relativistic electrons (in μOe), α is the spectral index at high frequencies ($\nu \gg \nu_m$), z is the redshift of the object, τ_0 is the optical depth at frequency ν_m, and $\Delta\theta$ is the angular size of the component in arcseconds. The optical depth τ_0 is determined by the condition for an extrema of $S(\nu)$ [20]:

$$e^{\tau_0} = 1 + \frac{2\alpha+5}{5}\tau_0. \qquad (4)$$

The frequency dependence of the radio spectral flux density is given by the expression [20]

$$S(\nu) = S_m \left(\frac{\nu}{\nu_m}\right)^{5/2} \frac{1-e^{-\tau_0\left(\frac{\nu}{\nu_m}\right)^{-\alpha-5/2}}}{1-e^{-\tau_0}}. \qquad (5)$$

We recalculated the spectra of the compact components of 3C 234 presented in Fig. 6 in accordance with (5), with $\nu_1 = 408$ MHz, $S_1 = 6.33$ Jy, $\alpha = 1$, $z = 0.185$, $H_{\perp} \approx 210$ μOe [10], and $\tau_0 = 0.64$ for the southwestern component 1 and the same values of α, z, τ_0, and ν_1 with $S_1 = 2.35$ Jy, and $H_{\perp} \approx 270$ μOe [10] for the northeastern component 5. The results, shown by the dashed curves, are in good agreement with the derived component flux densities for sizes of $0.55'' \pm 0.05''$ for the southwestern and $0.27'' \pm 0.03''$ for the northeastern component. These values are in good agreement with the direct measurements of [10] obtained using the VLA at 15 GHz, which show the sizes of these two components to be $0.7'' \times 0.3''$ and $0.4'' \times 0.25''$, respectively.

The integrated spectral index is 0.99 at frequencies above 1000 MHz and decreases to 0.85 at 1000–25 MHz. Supposing that this break is associated with variations in the spectrum of component 3, which is not observed at decameter wavelengths, we can derive this spectrum by taking the difference between the total flux density of the source and the flux densities of all the remaining components extrapolated to the low frequencies using the spectral indices determined at frequencies above 1000 MHz. The result is shown in Fig. 6 by the dotted curve *3*. The contribution of this component to the total flux density of the source becomes negligible below 400 MHz, and the spectral behavior of S_0 is determined by the sum of the flux densities of components 1, 5 shown in Fig. 6 and 2, 4 with each of these having a spectral index close to unity right down to about 50 MHz. The contribution of components 1 and 5 is reduced at lower frequencies due to self-absorption. The spectra of components 2 and 4 are not shown in Fig. 6, and we note only that their spectral indices, which are close to unity at centimeter and possibly longer (meter) wavelengths, increase at decameter wavelengths. As a result, these two components provide about 70% of the total flux density of 3C 234 at 20 and 25 MHz, as is noted above.

5. CONCLUSIONS

Our investigations of the angular structure of the radio galaxy 3C 234 using the URAN-1 and URAN-2 radio interferometers show that the coordinates of the hot spots in the extended lobes coincide with their positions at shorter wavelengths. It is obvious that the same is true of the positions and position angles of the elliptical Gaussians used to describe

the regions adjacent to the hot spots (components 2 and 4 in Fig. 5), although the sizes and relative flux densities of the lobes grow appreciably. The apparent sizes of the hot spots (taking into account scattering on electron-density inhomogeneities in the interstellar medium) are smaller than the resolution of the URAN-2 interferometer. However, our derived flux densities of these components indicate that there is self-absorption in the hot spots. On this basis, we derived the intrinsic sizes of these components, which are in good agreement with direct measurements at centimeter wavelengths. Note that features in the spectra of radio sources associated with the influence of plasma both in the source and along the line of sight are very common at such low frequencies. The analysis of such features provides unique opportunities to derive the intrinsic angular sizes of compact components together with the parameters of the plasma and magnetic fields in the source or the interstellar medium.

Our analysis shows that the extended, low-surface-brightness emission in the region of the core of the radio galaxy, which provides an appreciable fraction of the total flux density at centimeter wavelengths, is absent at decameter wavelengths. It is possible that this is, in fact, the origin of the change in the integrated spectral index below 1000 MHz. This tentative conclusion must be verified by direct measurements of the source structure at decimeter and meter wavelengths.

An extrapolation of the derived source structure and the spectra of its components indicates that the model radio-brightness distribution of 3C 234 will contain only two components at frequencies $\nu <$ 20 MHz, corresponding to the extended components 2 and 4. If the spectral indices of these components are preserved at lower frequencies, the integrated spectral index may again grow.

ACKNOWLEDGMENTS

This work was supported by the Russian Foundation for Basic Research (project no. 01-02-17811) and the Federal Target Science and Technology Program "Astronomy."

REFERENCES

1. P. Maltby and A. T. Moffet, Astrophys. J., Suppl. Ser. **7** (67), 141 (1962).
2. G. N. Macdonald, S. Kenderdine, and A. C. Neville, Mon. Not. R. Astron. Soc. **138**, 259 (1968).
3. P. N. Wilkinson, Mon. Not. R. Astron. Soc. **160**, 305 (1972).
4. F. M. Riley and G. G. Pooley, Mem. R. Astron. Soc. **80**, 105 (1975).
5. S. F. Burch, Mon. Not. R. Astron. Soc. **186**, 293 (1979).
6. S. F. Burch, Mon. Not. R. Astron. Soc. **186**, 519 (1979).
7. R. G. Strom and R. G. Conway, Astron. Astrophys., Suppl. Ser. **61**, 547 (1985).
8. P. Alexander, Mon. Not. R. Astron. Soc. **225**, 27 (1987).
9. M. J. Hardcastle, P. Alexander, G. G. Pooley, *et al.*, Mon. Not. R. Astron. Soc. **288**, 859 (1997).
10. J. W. Dreher, Astron. J. **86** (6), 833 (1981).
11. I. R. Allen, B. Anderson, R. G. Conway, *et al.*, Mon. Not. R. Astron. Soc. **124**, 477 (1962).
12. Yu. V. Volodin, A. G. Gubanov, and R. D. Dagkesamanskiĭ, Tr. Fiz. Inst. Akad. Nauk SSSR **189**, 173 (1988).
13. A. V. Megn, S. Ya. Braude, S. L. Rashkovskiĭ, *et al.,* Radiofiz. Radioastron. **2**, 385 (1997).
14. A. V. Megn, L. G. Sodin, N. K. Sharykin, *et al.*, in *Antennas,* Ed. by A. A. Pistol'kors (Svyaz', Moscow, 1978) [in Russian], No. 26, p. 15.
15. A. V. Megn, S. L. Rashkovskiĭ, V. A. Shepelev, *et al.,* Radiofiz. Radioastron. **3**, 284 (1998).
16. D. Himmelblau, *Process Analysis by Statistical Methods* (McGraw-Hill, New York, 1970; Mir, Moscow, 1973).
17. A. V. Megn, S. L. Rashkovskiĭ, and V. A. Shepelev, Radiofiz. Radioastron. **6**, 9 (2001).
18. H. Kuhr, U. Nauber, I. I. K. Pauliny-Toth, and A. Witzel, Preprint Max-Planck-Inst. Radioastron. (Bonn, 1981), p. 1.
19. N. H. Bedford, A. J. Kerr, S. N. Mathur, *et al.*, Mon. Not. R. Astron. Soc. **195**, 259 (1981).
20. S. Ya. Braude and A. V. Megn, Kinemat. Fiz. Neb. Tel **13**, 74 (1997).

Translated by D. Gabuzda

Astronomy Reports, Vol. 47, No. 12, 2003, pp. 1046–1053. Translated from Astronomicheskiĭ Zhurnal, Vol. 80, No. 12, 2003, pp. 1136–1143.
Original Russian Text Copyright © 2003 by Barsukov, Tsygan.

Returning Positron Flux in the Polar-Cap Regions of a Radio Pulsar

D. P. Barsukov and A. I. Tsygan

Ioffe Physicotechnical Institute, Russian Academy of Sciences, ul. Politekhnicheskaya 26, St. Petersburg, 194021 Russia

Received January 16, 2003; in final form, May 8, 2003

Abstract—An approximate method for calculating the returning positron flux in the polar-cap regions of a radio pulsar is proposed. The pulsar is considered in the Goldreich–Julian model for a regime of free-electron emission from the neutron-star surface in the region of open lines of the dipolar magnetic field. Calculations have been done for the case when the dipolar magnetic moment is aligned with the star's rotational axis. The acceleration of primary electrons is assumed to occur near the neutron-star surface on scales comparable to the transverse radius of the tube of open field lines. The generation of electron–positron pairs by curvature radiation of the primary electrons is taken into account. A considerable contribution to the returning flux is made by the region where the electric field is screened by the electron–positron plasma. © *2003 MAIK "Nauka/Interperiodica".*

1. INTRODUCTION

The problem of the magnitude of the returning positron flux in the polar-cap regions of open field lines has not been definitely solved. The magnitude of the returning flux in a pulsar flux tube was first estimated by Fawley *et al.* [1]. Pal'shin and Tsygan [2] later proposed a system of algebraic equations that could be used to find the returning flux in a pulsar for which the height of the upper plate (anode) of the pulsar tube diode was large. In 2001–2002, Harding and Muslimov [3, 4] considered in detail the effect of the generation of electron–positron pairs on pulsar electrodynamics, and estimated the magnitude of the returning flux assuming an exponential decrease of the longitudinal electric field. The electrodynamics of a pulsar tube taking into account the birth of secondary electron–positron pairs was studied in detail by Shibata *et al.* [5, 6] in a model with an "outer gap" (in the region close to the light cylinder).

In this paper, we present a model that can be used to find the magnitude of the returning flux in the polar-cap regions of radio pulsars. We take into account the contribution to the returning positron flux made by the region in which the electric field is screened by the electron–positron plasma. We compare our results with those of other studies.

2. ELECTRIC FIELD IN A FREE-EMISSION REGIME

We will use here the results of Muslimov and Tsygan [7], who constructed an electrodynamic model for a pulsar including general relativistic effects. They found the electric potential Φ using the Poisson equation

$$\mathrm{div}\left(\frac{1}{\alpha}\nabla\Phi\right) = -4\pi(\rho+\rho_{\mathrm{eff}}), \qquad (1)$$

where $\rho_{\mathrm{eff}} = \frac{1}{4\pi c}\mathrm{div}\left\{\frac{1}{\alpha}\left(1-\frac{k}{\eta^3}\right)[[\mathbf{\Omega}\times\mathbf{r}]\times\mathbf{B}]\right\}$ is the Goldreich–Julian charge density, ρ is the charge density of the particles escaping from the neutron-star surface, $\alpha\sqrt{1-\epsilon/\eta}$, $\epsilon = r_g/a$, $r_g = 2GM/c^2$ is the gravitational radius of the neutron star, r is the radial coordinate in the Schwarzschild metric, a is the neutron-star "radius," $\eta = r/a$, $k\epsilon\beta$, $\beta = I/I_0$ is the star's moment of inertia in units of $I_0 = Ma^2$, $\mathbf{\Omega}$ is the angular-velocity vector of the neutron star's rotation, and $\mathbf{B}$ is the magnetic-field intensity.

In the case of a dipolar field and in the absence of the generation of secondary particles, ρ and ρ_{eff} are

$$\rho = \frac{\Omega B_0}{2\pi c}\frac{1}{\alpha\eta^3}\frac{f(\eta)}{f(1)}\left[A(\xi)\cos\chi + \frac{3}{2}D(\xi)\sin\chi\cos\phi\right], \qquad (2)$$

$$\rho_{\mathrm{eff}} = \frac{\Omega B_0}{2\pi c}\frac{1}{\alpha\eta^3}\frac{f(\eta)}{f(1)}$$
$$\times\left[\left(1-\frac{k}{\eta^3}\right)\cos\chi + \frac{3}{2}H(\eta)\theta(\eta)\xi\sin\chi\cos\phi\right],$$

where $A(\xi)$ and $D(\xi)$ are functions of the variable ξ only, χ is the angle between the magnetic moment and the rotational axis, B_0 is the field at the magnetic pole, $\theta(\eta) = \theta_0\sqrt{\eta\frac{f(1)}{f(\eta)}}$, $\theta_0 = \arcsin\sqrt{\frac{a}{R_{\mathrm{LC}}f(1)}}$ is the magnetic latitude at the stellar surface of the last open

1063-7729/03/4712-1046$24.00 © 2003 MAIK "Nauka/Interperiodica"

magnetic line crossing the light cylinder, $R_{\rm LC} = c/\Omega$ is the radius of the light cylinder,

$$f(\eta) = -3\left(\frac{\eta}{\epsilon}\right)^3\left[\ln\left(1-\frac{\epsilon}{\eta}\right)+\frac{\epsilon}{\eta}\left(1+\frac{\epsilon}{2\eta}\right)\right],$$

$$H(\eta) = \frac{1}{\eta}\left(\epsilon-\frac{k}{\eta^2}\right)+\left(1-\frac{3}{2}\frac{\epsilon}{\eta}+\frac{1}{2}\frac{k}{\eta^3}\right)\times\left[f(\eta)\left(1-\frac{\epsilon}{\eta}\right)\right]^{-1},$$

and the Poisson equation is

$$\frac{1}{\eta^2}\frac{\partial}{\partial\eta}\left(\eta^2\frac{\partial\Phi}{\partial\eta}\right)+\frac{1}{\alpha^2\eta^2\theta^2}\left[\frac{1}{\xi}\frac{\partial}{\partial\xi}\left(\xi\frac{\partial\Phi}{\partial\xi}\right)+\frac{1}{\xi^2}\frac{\partial^2\Phi}{\partial\phi^2}\right] = -4\pi\frac{\Omega B_0}{2\pi c}a^2\frac{1}{\alpha^2\eta^3}\frac{f(\eta)}{f(1)}\times\left\{\left[1-\frac{k}{\eta^3}+A(\xi)\right]\cos\chi+\frac{3}{2}[\xi\theta(\eta)H(\eta)+D(\xi)]\sin\chi\cos\phi\right\}. \quad (3)$$

Here and below, we use the corotating system of coordinates (t, η, ξ, ϕ) introduced in [7], where t is the time coordinate, the coordinates $x^1 \equiv \eta$ and ξ vary along and across the field lines ($\xi = 0$ corresponds to the flux-tube center, and $\xi = 1$ to the last open magnetic line), and ϕ is the azimuthal angle.

In accordance with the Goldreich–Julian model and with the free emission of charges from the surface, the following boundary conditions are imposed on the potential Φ:

$$\Phi|_{\eta=1} = 0,\ \left.\frac{\partial\Phi}{\partial\eta}\right|_{\eta=1} = 0, \quad (4)$$

$$\Phi\,|_{\xi=1} = 0,\ \ \Phi|_{\xi=0}\ \ \text{bounded}.$$

Tsygan [8] considered the electrodynamics of a pulsar-tube diode in the presence of an upper plate (anode) at height $z_c \equiv \eta_c - 1$. By definition, it is assumed that at this height

$$E_{||} \equiv -\frac{1}{a}\frac{\partial\Phi}{\partial\eta} = 0. \quad (5)$$

In particular, Tsygan [8] gave the solution of (3) for the case $z_c \ll 1$:

$$\Phi(z,\xi,\phi) = \Phi_0\left[\frac{6k}{\alpha^2}\cos\chi\sum_{i=1}^{\infty}\frac{2}{k_iJ_1(k_i)}\times F(z,\gamma_i)J_0(\tilde{k}_i\xi)+\frac{3\theta_0}{\alpha^2}H(1)\delta(1)\sin\chi\cos\phi\times\sum_{i=1}^{\infty}\frac{2}{\tilde{k}_iJ_2(k_i)}F(z,\tilde{\gamma}_i)J_1(\tilde{k}_i\xi)\right], \quad (6)$$

where

$$F(z,\gamma_i) = \frac{1}{\gamma_i^2}\left[\frac{1}{2}\epsilon_i(e^{\gamma_i z}+e^{-\gamma_i z}-2)+z-\frac{1}{\gamma_i}(1-e^{-\gamma_i z}),\right],$$

$$\epsilon_i = \frac{2}{\gamma_i}\frac{\gamma_i z_c+(\gamma_i+1)(e^{-\gamma_i z_c}-1)}{(\gamma_i-1)(e^{\gamma_i z_c}-1)-(\gamma_i+1)(e^{-\gamma_i z_c}-1)},$$

$\Phi_0 = (\Omega a/c)B_0 a$, $\gamma_i = k_i/\theta_0\sqrt{1-\epsilon}$, $\tilde{\gamma}_i = \tilde{k}_i/\theta_0\times\sqrt{1-\epsilon}$, k_i and $\tilde{k}_i$ are the roots of the equations $J_0(k_i) = 0$ and $J_1(\tilde{k}_i) = 0$, respectively, and $z \equiv \eta - 1$ everywhere.

In the case $\theta_0 \sim z_c \ll 1$, we can rewrite (6) for an axisymmetric pulsar ($\chi = 0$) as

$$\Phi(z,\xi) = \Phi_0 6k\theta_0^2\left[\frac{1}{4}z(1-\xi^2)-\sqrt{1-\epsilon}\theta_0\sum_{i=1}^{\infty}\tilde{\Phi}_i(z)J_0(k_i\xi)\right], \quad (7)$$

where

$$\tilde{\Phi}_i = \frac{2}{k_i^4J_1(k_i)}\frac{1-e^{-\gamma_i z}}{1+e^{-\gamma_i z_c}}\left[1+e^{-\gamma_i(z_c-z)}\right]. \quad (8)$$

If $z_c \ll \theta_0$, the solution (6) becomes

$$\Phi = \frac{\Phi_0}{1-\epsilon}z^2\left(\frac{z_c}{2}-\frac{z}{3}\right)\left[3k\cos\chi+\frac{3}{2}\xi\theta_0H(1)\delta(1)\sin\chi\cos\phi\right], \quad (9)$$

which corresponds with the solution of Beskin [9] for $\chi = 0$ to within the factor $1/(1-\epsilon)f(1)$.

Pal'shin and Tsygan [2] proposed the following system of equations to find the returning flux for long, thin tubes:

$$\rho_{\rm eff}(1)+\rho_e(1)+\rho_+(1) = 0,$$
$$\rho_{\rm eff}(\eta_c)+\rho_e(\eta_c)+\rho_-(\eta_c) = 0, \quad (10)$$

where ρ_+ is the charge density associated with the returning flux of positrons, $\rho_- = -\rho_+$ is the charge density created by secondary electrons that has become uncompensated due to the return of some fraction of the positrons, and ρ_e is the charge density created by the primary electrons. In [2], this system of equations was derived from the condition that the longitudinal electric fields at the neutron-star surface and at the upper plate of the diode be approximately equal to zero ($\eta = \eta_c$). In particular, for an aligned pulsar with a dipolar magnetic field, the solution of these equations becomes [2]

$$\frac{\rho_+(1)}{\rho_{\rm eff}(1)} = \frac{k}{2}\frac{1}{1-k}\left(1-\frac{1}{\eta_c^3}\right). \quad (11)$$

3. MAIN ASSUMPTIONS

We used the following model for the pulsar flux tube.

(1) The tube is treated in the framework of the Goldreich–Julian model, and the diode is located at the neutron star's surface (a polar-cap model).

(2) We assume that the tube is axisymmetric (in particular, aligned, $\chi = 0$) and stationary.

(3) The pulsar operates in a regime with free electron emission from the neutron-star surface.

In accordance with the diode structure proposed by Arons and Scharlemann [10], we choose the current of the primary electrons such that the longitudinal electric field vanished at a height of $z_c = \eta_c - 1$. In the absence of generation above this boundary, the longitudinal field changes its sign, and the potential begins to decrease to zero and lower. We choose the height z_c so that the birth of pairs slightly lower occurs at a rate that is sufficient for the resulting negative charge due to the returning positrons to maintain the potential growth, so that the field itself would tend to zero.

We believe that a different choice of z_c is probably not compatible with the condition of stationarity. If we choose a value of z_c that was too large, the mass production of electron–positron pairs would begin in the region of strong electric field. This would result in the return of more than one positron per primary electron, blocking the emission of primary electrons from the neutron-star surface; essentially, the pulsar flux tube would then no longer be in a stationary state. On the other hand, if the value of z_c is too low, the number of positrons born will be insufficient to support the electric field via their return; the field above z_c will change its sign, and the electric potential Φ will begin to decrease. We suppose that this mode will likewise not be stationary, because low-energy secondary electrons will accumulate in the resulting potential well.

As is seen from the results obtained, the electric field decreases approximately exponentially nearly to zero a small distance above z_c; this is directly associated with the rapid growth of the rate of generation of electron–positron pairs by curvature-radiation photons, and coincides fully with the results of Harding and Muslimov [3]. However, since $z_c \sim 0.02$–0.03 for the pulsars considered, we should make the following qualification. The point is that the main reason the electric field drops to zero and below is the positive charge provided by ρ_{eff}. For such low diodes, an increase of ρ_{eff} at the height z_c already provides a reversal of the electric-field sign and the onset of the decrease in the electric potential. On the other hand, the main change of ρ_{eff} takes place between $z = 0.1$ and $z = 1$ (at greater heights for orthogonal pulsars, up to the light cylinder). It is obvious that the electric field must be maintained in this region, as well.

If the electric field is maintained in this region (as at low heights) by the returning positron flux, then, in spite of its extreme weakness, the field will also be nonzero in this region, requiring the return of about $k/2$ positrons per primary electron. Hence, the main provider of the returning flux is precisely the zone in which the electric field is virtually completely screened.

In this connection, we assume that the growth of ρ_{eff} is compensated by the return of secondary positrons up to heights of the order of $\eta_f = \sqrt[4]{2k/\sin^2\theta_0}$.

4. ELECTRIC FIELD OF THE DIODE TAKING INTO ACCOUNT PARTICLE GENERATION

Let us now take into account the effect of the generation of electron–positron pairs on the electrodynamics of the pulsar tube. We will use the approach described in [1]; a similar approach was used in [6].

First, we note that, in the zone where there is no generation of secondary particles and the positrons in the returning flux have become relativistic, there will be no changes in Eq. (3) except for the constant A, which will now describe the total flux of electrons and positrons. It is obvious that the condition of the free emission of charges from the surface $E_{\parallel} = 0$ will also remain unchanged (if the number of returning positrons per primary electron is less than unity, then we can always provide an excess negative charge near the surface by increasing the supply of electrons from the stellar surface, which is necessary for the existence of the electron flux from the neutron-star surface in the free charge-emission regime).

Let $q(x)$ positrons and electrons be born per unit volume per unit time with the spectrum $df/d\Gamma$, where $\Gamma = 1/\sqrt{1 - v_{\parallel}^2/c^2}$ is the gamma factor of the longitudinal motion of a particle ($v_{\parallel}$ is the velocity of the particle along the magnetic field). Because the magnetic field is strong, particles born on a given magnetic-field line will remain on that line. We assume that the electron–positron pairs are born moving upward, away from the star. In particular, this is the case for pairs born by curvature-radiation photons generated by the primary electrons. After their birth, the upward-moving positrons gradually decelerate, and some fraction of them stop and begin to move downward after having reached a certain height. In contrast to the positrons, the electrons will accelerate and always escape to infinity.

In this case, the total electric-charge density created by the secondary electrons and positrons will be

$$\rho_s = \sqrt{\gamma_{11}}\tilde{\rho} + (\sqrt{\gamma_{11}} - \sqrt{h}g_1)\frac{1}{\sqrt{-g}}\rho_0, \qquad (12)$$

where

$$\tilde{\rho} = \int_1^\eta d\eta_y \sqrt{\frac{\gamma(y)}{\gamma(x)}} q(y) \int_1^\infty \frac{df}{d\Gamma}(y,\Gamma_y) \qquad (13)$$
$$\times \left(\frac{\Gamma}{\sqrt{\Gamma^2-1}} - \frac{\Gamma_-}{\sqrt{\Gamma_-^2-1}}\right) d\Gamma$$
$$+ \int_1^\eta d\eta_y \sqrt{\frac{\gamma(y)}{\gamma(x)}} q(y) \int_1^{\frac{1+U_\infty - U(x)}{\sqrt{h}}} \frac{df}{d\Gamma}(y,\Gamma_y)$$
$$\times \left(\frac{\Gamma}{\sqrt{\Gamma^2-1}} - 1\right) d\Gamma - \int_1^\eta d\eta_y \sqrt{\frac{-g(y)}{-g(x)}} q(y)$$
$$\times \int_1^{\frac{\sqrt{h}+U(x)-U(y)}{\sqrt{h_y}}} \frac{df}{d\Gamma}(y,\Gamma_y) \left(\frac{\Gamma_-}{\sqrt{\Gamma_-^2-1}} + 1\right) d\Gamma_y$$
$$+ \int_\eta^{+\infty} d\eta_y \sqrt{\frac{-g(y)}{-g(x)}} q(y) \int_1^{\frac{1+U_\infty - U(y)}{\sqrt{h_y}}} \frac{df}{d\Gamma}(y,\Gamma_y)$$
$$\times \left(\frac{\Gamma_+}{\sqrt{\Gamma_+^2-1}} - 1\right) d\Gamma_y,$$

$$\rho_0 = \int_1^{+\infty} d\eta_y \sqrt{-g(y)} q(y) \int_1^{\frac{1+U_\infty - U(y)}{\sqrt{h_y}}} \frac{df}{d\Gamma}(y,\Gamma_y) d\Gamma_y.$$

In these formulas,

$$\Gamma_- = \Gamma_+ + \frac{2}{\sqrt{h(x)}}(U(x) - U(y))$$
$$= \sqrt{\frac{h(y)}{h(x)}}\Gamma_y + \frac{U(x)-U(y)}{\sqrt{h(x)}},$$
$$\Gamma_y(\Gamma) = \sqrt{\frac{h(x)}{h(y)}}\Gamma + \frac{U(x)-U(y)}{\sqrt{h(y)}},$$
$$\Gamma_+(\Gamma_y) = \sqrt{\frac{h(y)}{h(x)}}\Gamma_y - \frac{U(x)-U(y)}{\sqrt{h(x)}},$$

$U = e\Phi/mc^2$, $U_\infty = e\Phi_\infty/mc^2$ is the potential on a given magnetic-field line as $\eta \to +\infty$, g_{ik} is the metric space–time tensor, $\gamma_{\alpha\beta}$ is the space metric, $h = g_{00}$, $g = \det(g_{ik})$, $\gamma = \det(\gamma_{\alpha\beta})$, and $g_\alpha = -g_{0\alpha}/g_{00}$. We have used the notation $y = (t_y, \eta_y, \xi_y, \phi_y)$, $x = (t, \eta, \xi, \phi)$ and integrated along the curves ($t_y = t$, $\xi_y = \xi$, $\phi_y = \phi$), which by definition coincide with the field lines.

Note that the ρ_0 term describes the flux of relativistic positrons arriving from infinity, which, below the the generation zone, coincides with the returning positron flux. In this case, since $g_1 = 0$ for an axisymmetric pulsar in the given coordinate system, there are no differences in the η dependence of the charge density for a relativistic flux from the stellar surface and a relativistic flux from infinity to the pulsar surface. Thus, the additional term with ρ_0 simply renormalizes the flux of primary electrons.

As a result, the Poisson equation becomes

$$\mathrm{div}\left(\frac{1}{\alpha}\nabla\Phi\right) = -4\pi(\rho_s + \rho_e + \rho_{\mathrm{eff}}), \qquad (14)$$

where ρ_e is the charge density due to the primary electrons.

We use the conditions (4) as boundary conditions, also imposing on the potential Φ the condition

$$\Phi \to \Phi_\infty \qquad \text{for} \qquad \eta \to +\infty. \qquad (15)$$

This last condition was proposed in [3]; it ensures that the solution will automatically satisfy the requirement $E_\parallel \to 0$ for $\eta \to +\infty$.

Let us make some simplifying assumptions.

(1) We neglect cross terms in the Laplacian (which arise from the non-orthogonality of the chosen coordinate system).

(2) We assume that the secondary electrons move at the speed of light everywhere, and that the secondary positrons before and after the turning point also move at the speed of light.

(3) We assume for simplicity that the spectrum of the born particles $df/d\Gamma$ is constant along a magnetic-field line; we adopt for this spectrum the actual spectrum generated by the curvature-radiation photons at height $1.5z_c$.

(4) We neglect the effect of the gravitational redshift on the energy of the electrons and positrons.

(5) Since the relative increase of the primary flux due to the returning positron flux is a small quantity of the order of $k \sim 0.15$, we neglect the change of the rate of generation of electron–positron pairs, since this correction will be insignificant near z_c owing to the rapid growth of $q(x)$, and, at greater heights, the value of $q(x)$ affects the solution only weakly.

The equations then become

$$\frac{1}{\eta^2}\frac{\partial}{\partial\eta}\left(\eta^2\frac{\partial\Phi}{\partial\eta}\right) + \frac{1}{\alpha^2\eta^2\theta^2}\frac{1}{\xi}\frac{\partial}{\partial\xi}\left(\xi\frac{\partial\Phi}{\partial\xi}\right) \qquad (16)$$

$$= -4\pi\frac{\Omega B_0}{2\pi c}\frac{1}{\alpha^2\eta^3}\frac{f(\eta)}{f(1)}\left(A(\xi)+I(U)+1-\frac{k}{\eta^3}\right),$$

$$I(U) = -2\int_1^{\eta} d\eta_y\sqrt{\gamma(y)}q(y)\int_1^{1+U(x)-U(y)}\frac{df}{d\Gamma}(\Gamma_y)d\Gamma_y.$$

Since we consider below only rapidly rotating pulsars with strong magnetic fields, then, in accordance with the results of [3, 4, 11], we can neglect the generation of electron–positron pairs by photons generated during the inverse Compton scattering of thermal radiation from the neutron-star surface on the primary electrons. In this connection, $q(x)$ denotes the rate of generation of electron–positron pairs from curvature-radiation photons.

5. SOLUTION OF THE EQUATIONS

Let us divide the flux tube into three parts:

(1) the diode zone (the multiplication factor is less than 0.001);

(2) the transition layer;

(3) the screening zone (the multiplication factor is greater than 1).

In the diode zone, the rate of particle generation is extremely low, and we will accordingly assume that, in this zone, the field at a given ξ coincides with the solution (7) for a vacuum diode at some height z_c; each ξ will have its own height z_c. Note that this assumption is automatically fulfilled when we can neglect the effect of the transverse part of the Laplacian on the potential. Below, we will designate the solution (7) continued to the region $z \geq z_c$ by a constant potential along the field lines $\Phi_0(z,\xi)$:

$$\Phi_0(z,\xi) = \Phi(z,\xi) \quad \text{for} \quad z \leq z_c, \qquad (17)$$
$$\Phi_0(z,\xi) = \Phi(z_c,\xi) \quad \text{for} \quad z \geq z_c,$$

where $\Phi(z,\xi)$ is the solution of (7).

With regard to the transition layer and screening zone, we note that the total potential drop along a field line is only a few percent of the potential drop in the diode zone. This means that we can use the transverse part of the Laplacian of the potential Φ_0 in place of the transverse part of the Laplacian of Φ (moreover, for diodes at low heights, this is small compared to $\partial^2\Phi/\partial\eta^2$). Thus, the problem becomes one-dimensional. Moreover, in the screening zone, the electric field along the field lines becomes so small that we can treat the longitudinal part of the Laplacian $\frac{1}{\eta^2}\frac{\partial}{\partial\eta}\eta^2\frac{\partial\Phi}{\partial\eta}$ as a small perturbation compared to the large terms in the right-hand side of (16).

We set the natural conditions of the continuity of the potential Φ and of its derivative with respect to η at the zone boundaries, and assume the continuity of the electric charge density.

Thus, the problem reduces to the following:

—in the diode zone, to using the already known analytical solution $\Phi_0(z,\xi)$ with the free parameter z_c;

—in the transition layer, to the solution of the boundary-value problem for the one-dimensional equation

$$\frac{1}{\eta^2}\frac{\partial}{\partial\eta}\left(\eta^2\frac{\partial\Phi}{\partial\eta}\right)+\frac{1}{\alpha^2\eta^2\theta^2}\frac{1}{\xi}\frac{\partial}{\partial\xi}\left(\xi\frac{\partial\Phi_0}{\partial\xi}\right)$$
$$= -4\pi\frac{\Omega B_0}{2\pi c}\frac{1}{\alpha^2\eta^3}\frac{f(\eta)}{f(1)}\left(A(\xi)+I(U)+1-\frac{k}{\eta^3}\right);$$

—in the screening zone, to the solution of the algebraic equation with respect to $U = e\Phi/mc^2$ at each point

$$-4\pi\frac{\Omega B_0}{2\pi c}\frac{1}{\alpha^2\eta^3}\frac{f(\eta)}{f(1)}\left(A(\xi)+I(U(\eta))\right. \qquad (18)$$
$$\left.+1-\frac{k}{\eta^3}\right) = \frac{1}{\alpha^2\eta^2\theta^2}\frac{1}{\xi}\frac{\partial}{\partial\xi}\left(\xi\frac{\partial\Phi_0}{\partial\xi}\right).$$

Number density of positrons in the returning flux, height of the upper diode plate, and temperature at the neutron-star surface

ξ	n_+/n_B, 10^{-2}	z_c, 10^{-2}	T, 10^6 K
	$B_0 = 0.1B_{\rm cr}, P = 100$ ms		
0.3	7.00	3.6	7.2
0.5	7.03	3.3	6.7
0.7	6.97	3.4	6.2
	$B_0 = 0.5B_{\rm cr}, P = 300$ ms		
0.3	6.96	3.4	7.8
0.5	7.0	3.1	7.3
0.7	6.93	3.2	6.9
	$B_0 = 0.09B_{\rm cr}, P = 33$ ms		
0.3	7.25	2.2	9.0
0.5	7.27	2.0	8.3
0.7	7.27	1.9	7.8

6. RESULTS

The table lists the obtained number densities of positrons of the returning current at the neutron-star surface n_+ in units of $n_B = \frac{\Omega B_0}{2\pi ec}\frac{1}{\alpha}$ and the heights of the diode upper plate z_c for some values of the field B_0 (in units of $B_{\rm cr} = 4.41\times10^{13}$ G) and pulsar

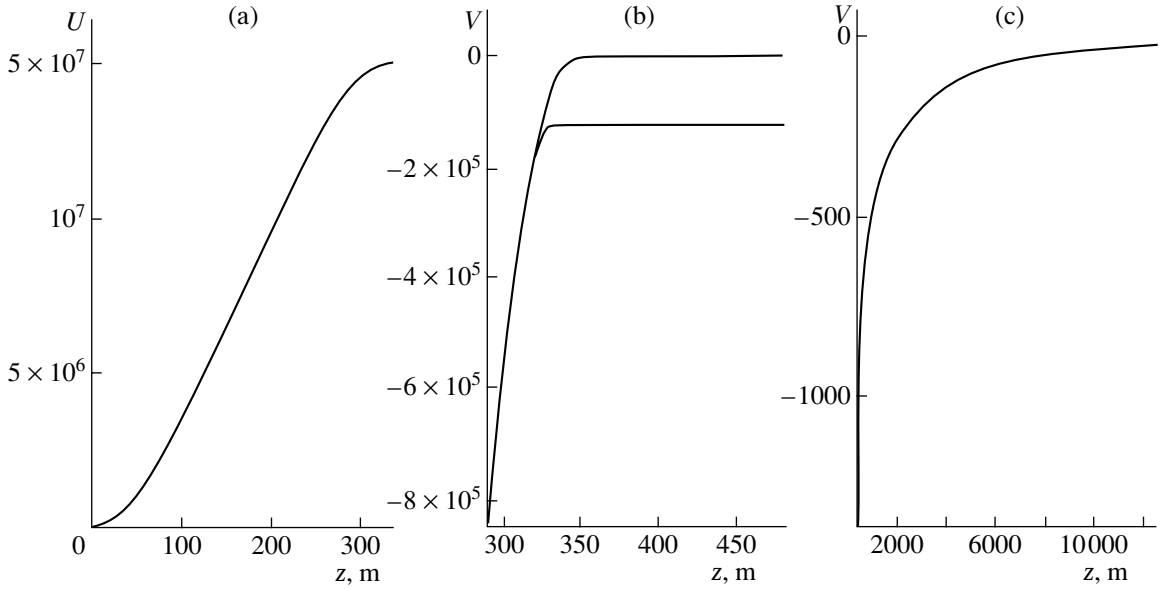

Fig. 1. Potential $\Phi(z,\xi)$ versus z in units of $e\Phi/mc^2$ for $\xi = 0.5$, $B_0 = 0.1B_{\rm cr}$, and $P = 100$ ms. The potential is measured from its value at infinity in panels (b) and (c) and from the potential at the stellar surface in panel (a). The lower curves in (a) and (b) represent the potential $\Phi_0(z,\xi)$.

period P. Everywhere we used the value $k = 0.15$ when obtaining these results. The temperature T of the neutron-star surface in regions of open field lines was estimated from the formula $\sigma T^4 = n_+ ce\Phi(z_c,\xi)$, where σ is the Stefan–Boltzmann constant.

Figures 1 and 2 show the resulting potential $\Phi(z,\xi)$ as a function of z for $\xi = 0.5$, $B_0 = 0.1B_{\rm cr}$, and $P = 100$ ms, and for $\xi = 0.5$, $B_0 = 0.5B_{\rm cr}$, and $P = 300$ ms. The regions shown in Figs. 1b and 2b correspond approximately to the transition layer and to the beginning of the screening zone; Figs. 1c and 2c present the variations of the potential Φ in the bulk of the screening region, where virtually all of the returning flux is formed. Figure 3 shows the secondary-particle spectrum $\dfrac{df}{d\Gamma}(\Gamma)$ we used to calculate the run of the potential for $\xi = 0.5$, $B_0 = 0.1B_{\rm cr}$, and $P = 100$ ms.

These results confirm that the system of equations (10) can also be used to find the returning flux in pulsars in which the height of the upper plate is low. However, the value of η_c must then be set equal to $+\infty$, instead of the height of the onset of the screening zone. The differences of these equations from the solution (11) are mainly due to the difference of $\rho + \rho_{\rm eff}$ from zero at the neutron-star surface and our neglect of the perpendicular component of the Laplacian $\dfrac{1}{a^2\eta^2\theta(\eta)^2}\dfrac{1}{\xi}\dfrac{\partial}{\partial\xi}\xi\dfrac{\partial\Phi}{\partial\xi}$ in the second of these equations.

We found that, with the assumptions made, the returning flux is virtually independent of the spectrum of $\dfrac{df}{d\Gamma}$ and the pair-generation rate $q(x)$ in the screening zone. Thus, the returning current is determined mainly by the parameters in the transition layer. In this case, the screening-zone potential difference that gives rise to the returning flux depends strongly on the shape of the spectrum and, as is fairly obvious, is equal to order of magnitude to the gamma factor of the secondary particles at the peak of their spectrum.

Concerning our neglect of the contribution to the density ρ_s made by nonrelativistic secondary positrons, we note that, to give birth to a pair in a strong magnetic field, it is sufficient for a photon to acquire a small angle Ψ between the wave vector $\mathbf{k}$ and the vector $\mathbf{B}$; this results in the birth of positrons with large longitudinal gamma factors $\Gamma \sim 1/\Psi$. In this connection, the velocity of the positrons born in the screening zone virtually coincides with the speed of light. Taking into account the nonrelativistic secondary positrons born in the transition layer (it is precisely these that make up the bulk of the returning flux) at heights above the point of their birth for the case of sufficiently high and smooth spectra affect the result very weakly; this agrees with the result obtained by Harding and Muslimov [3].

In the general case, the nonrelativistic nature of the positrons results in the appearance of an additional positive charge, which must also be compensated by the creation of a negative charge via the return of positrons. As a result, taking this into account in the stationary case can only lead to an increase of the returning flux.

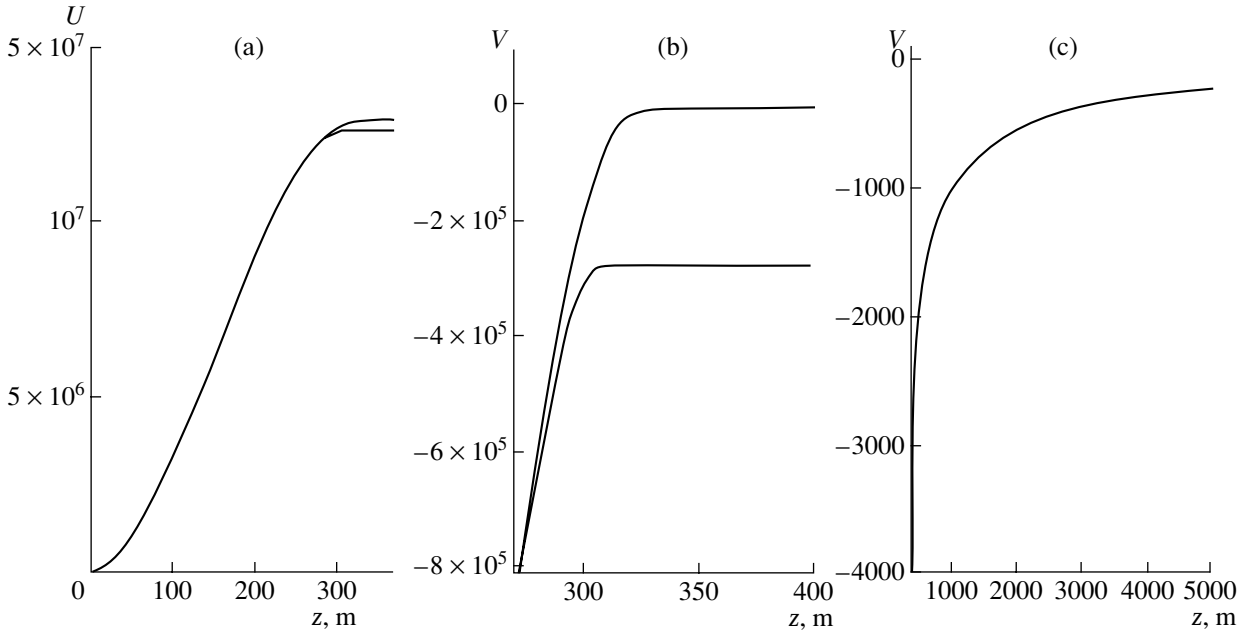

Fig. 2. Same as Fig. 1 for $\xi = 0.5$, $B_0 = 0.5B_{cr}$, and $P = 300$ ms.

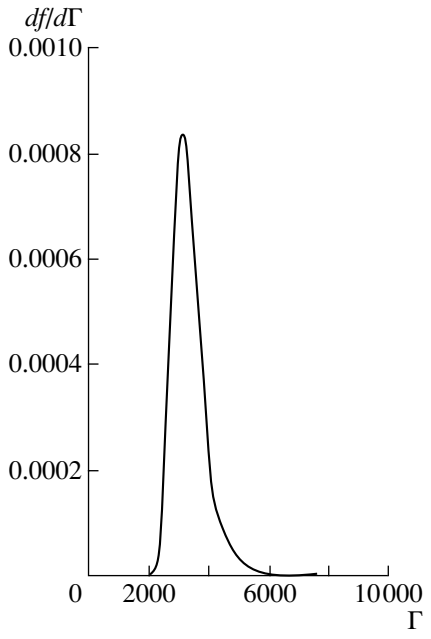

Fig. 3. Spectrum of the secondary particles $\frac{df}{d\Gamma}(\Gamma)$ for $\xi = 0.5$, $B_0 = 0.1B_{cr}$, and $P = 100$ ms.

When analyzing the results obtained, we found that the regime of large values of $\eta \geq \eta_f$, where $\eta_f = \sqrt{4}2k/\sin^2\theta_0$ for a dipolar magnetic field, requires more detailed study. At these heights, $\frac{\partial\rho_{\text{eff}}}{\partial\eta}$ changes its sign and ρ_{eff} begins to supply an additional negative charge, which must also be compensated to keep $E_\parallel$ small. We believe that, in this region, the compensation can be achieved only by the presence of a medium of positrons at rest; the appearance at lower heights of a medium of electrons at rest seems to us improbable. (With the presence of a medium near the transition layer, the latter would result in returning fluxes close to those obtained in [1, 6].) The point is that we could not obtain a solution satisfying the necessary boundary conditions at a point below η_f assuming that there are no discontinuities in the charge density (in particular, we assume the continuity of $\frac{\partial^2\Phi}{\partial\eta^2}$) and that the force acting on this medium from the beam of secondary particles is not too great [5].

We conclude that the points where ρ_{eff} reaches its maximum or minimum along field lines are important for the electrodynamics of the pulsar. In particular, the presence of these points enables us to obtain solutions with the boundary conditions (4) and the additional conditions $E_\parallel = 0$ and $\frac{\partial^2\Phi}{\partial\eta^2} = 0$ at a certain height.

7. CONCLUSION

We have presented here a model that can be used to derive the returning positron flux for an aligned pulsar with an axially symmetric magnetic field; specific results are given only for a dipolar field. In this connection, we can hardly expect these results to

coincide with the magnitudes of the returning fluxes in real pulsars.

However, the proposed model demonstrates that, for pulsars in which the height of the upper plate is small, including the contribution of the screening region to the returning flux increases this flux by a factor of approximately 10 over previous estimates, in particular, compared to the results of [3].

In the proposed model, we should not expect a considerable decrease of the returning flux when considering unaligned pulsars. In particular, for an orthogonal pulsar ($\chi = \pi/2$), a considerable contribution to the returning flux will be made by regions close to the light cylinder. In addition, the light cylinder will probably exceed in size the regions of the returning flux for an aligned pulsar (this is readily derived from (10) with $\chi = \pi/2$ and values of $a\eta_c$ near the light cylinder). Here, it is possible for the returning flux of secondary positrons to be very close to the magnitude of the flux of primary electrons, leading to a manifold increase of the latter (this is necessary to support the charge imbalance near the neutron-star surface).

Note that the model we have considered is not consistent with the observations for the Vela pulsar (in particular, the observed temperature of the polar caps is lower than we have obtained; see the Table). This probably implies that either this pulsar has an appreciably non-dipolar field, or $\mathbf{\Omega} \cdot \mathbf{B} < 0$, or that it operates in an outer-gap regime.

ACKNOWLEDGMENTS

This work was supported by the Russian Foundation for Basic Research (project no. 01-02-17811) and by the State Science and Technology Program in Astronomy.

REFERENCES

1. W. M. Fawley, J. Arons, and E. T. Scharlemann, Astrophys. J. **217**, 227 (1977).
2. V. L. Pal'shin and A. I. Tsygan, Preprint No. 1718 (Phys. Tech. Inst., St. Petersburg, 1998).
3. A. K. Harding and A. G. Muslimov, Astrophys. J. **556**, 987 (2001).
4. A. K. Harding and A. G. Muslimov, Astrophys. J. **568**, 862 (2002).
5. S. Shibata, J. Miyazaki, and F. Takahara, Mon. Not. R. Astron. Soc. **295**, L53 (1998).
6. S. Shibata, J. Miyazaki, and F. Takahara, Mon. Not. R. Astron. Soc. **336**, 233 (2002).
7. A. G. Muslimov and A. I. Tsygan, Mon. Not. R. Astron. Soc. **255**, 61 (1992).
8. A. I. Tsygan, Pis'ma Astron. Zh. **19**, 665 (1993).
9. V. S. Beskin, Pis'ma Astron. Zh. **16**, 665 (1990) [Sov. Astron. Lett. **16**, 286 (1990)].
10. J. Arons and E. T. Scharlemann, Astrophys. J. **231**, 854 (1979).
11. A. K. Harding and A. G. Muslimov, Astrophys. J. **508**, 328 (1998).
12. E. T. Scharlemann, J. Arons, and W. M. Fawley, Astrophys. J. **222**, 297 (1978).

Translated by G. Rudnitskiĭ

Astronomy Reports, Vol. 47, No. 12, 2003, pp. 1054–1062. Translated from Astronomicheskiĭ Zhurnal, Vol. 80, No. 12, 2003, pp. 1144–1158.
Original Russian Text Copyright © 2003 by Kocharov, Ogurtsov, Tsereteli.

Cosmogenic Radiocarbon as a Means of Studying Solar Activity in the Past

G. E. Kocharov[1], M. G. Ogurtsov[1], and S. L. Tsereteli[2]

[1] *Ioffe Physicotechnical Institute, Russian Academy of Sciences, ul. Politekhnicheskaya 26, St. Petersburg, 194021 Russia*

[2] *Tbilisi State University, pr. Chavchavadze 1, Tbilisi, 380028 Georgia*

Received March 12, 2003; in final form, July 26, 2003

Abstract—A series of yearly data on the concentration of radioactive carbon ^{14}C in tree rings measured at the Tbilisi State University in 1983–1986 and covering the time interval 1600–1940 is statistically analyzed. We find evidence for a 22-year cyclicity in the intensity of Galactic cosmic rays (GCRs) during the Maunder minimum of the solar activity (1645–1715), testifying that the solar dynamo mechanism continued to operate during this epoch. Variations of $\Delta^{14}C$ on timescales of tens and hundreds of years correlate well with the corresponding variations of the GCR intensity and solar activity, making radiocarbon a reliable source of information on long-timescale variations of solar activity in the past. Short-timescale (<30 years) fluctuations of $\Delta^{14}C$ may be appreciably distorted by time variations not associated directly with solar activity; probable origins of this distortion are discussed. © *2003 MAIK "Nauka/Interperiodica"*.

The analysis of solar activity (SA), its cyclicities, and its terrestrial manifestations long attracted the attention of scientists. The Sun is the main source of energy in the solar system and the generator of many processes in interplanetary space and on the Earth. Therefore, multi-faceted studies of the nature of solar activity are not only of great theoretical interest, but are also important for purely applied problems. However, information about processes on the Sun obtained from direct measurements provide information only about the current values of various SA parameters, and reveal variations in these parameters only on timescales comparable to the interval covered by the observations. In the case of sunspot numbers (Wolf numbers), this interval covers three centuries, whereas, for most other SA parameters, it covers only the last few decades.

At the same time, it is necessary to know the behavior of the Sun on timescales of hundreds and thousands of years for many practical purposes—forecasting solar activity in the future, analyzing the effect of solar activity on the climate of our planet, etc. For such purposes, we must use indirect sources of information about solar activity; one of these is cosmogenic radiocarbon, ^{14}C. Cosmogenic ^{14}C is produced in the Earth's atmosphere (mostly in the stratosphere) by energetic Galactic cosmic rays (GCRs), which are efficiently modulated by solar activity; it is then oxidized to CO_2, which takes part in a number of geophysical and geochemical processes (the global carbon-exchange cycle), and is ultimately stored in tree rings [1–3]. Thus, the variations of the radiocarbon concentration in tree rings can reflect either variations of the ^{14}C production rate in the atmosphere, i.e., variations in solar activity and the geomagnetic field, or changes in the character of the carbon cycle (climatic variations).

Studying solar activity over long timescales on the basis of the ^{14}C content in tree rings first began in the USSR in 1965–1967, as part of the program on astrophysical phenomena and radiocarbon studies formulated by Konstantinov and Kocharov [4]. The first long series of annual data on the radiocarbon concentration in tree rings was obtained at the Tbilisi State University (TSU) in 1983–1986 [5]. This data series covers the years 1600–1940; data for 1600–1700 were obtained from wood samples from the West Ukraine ($\approx 48°$ N), and for 1800–1940 from Karelian samples ($\approx 62°$ N). One of the main advantages of the Tbilisi series is that it covers the Maunder minimum (MM) of solar activity (1645–1715). This period was an epoch of considerable attenuation of solar activity, accompanied by an almost complete absence of sunspots. In-depth and comprehensive analyses of such abnormal periods, in particular, to elucidate whether or not the short-timescale (11–22 years) periodicity in solar activity persisted during these epochs, are important for the clarification of the nature of global SA minima and of mechanisms for solar cyclicity and sunspot formation.

1063-7729/03/4712-1054$24.00 © 2003 MAIK "Nauka/Interperiodica"

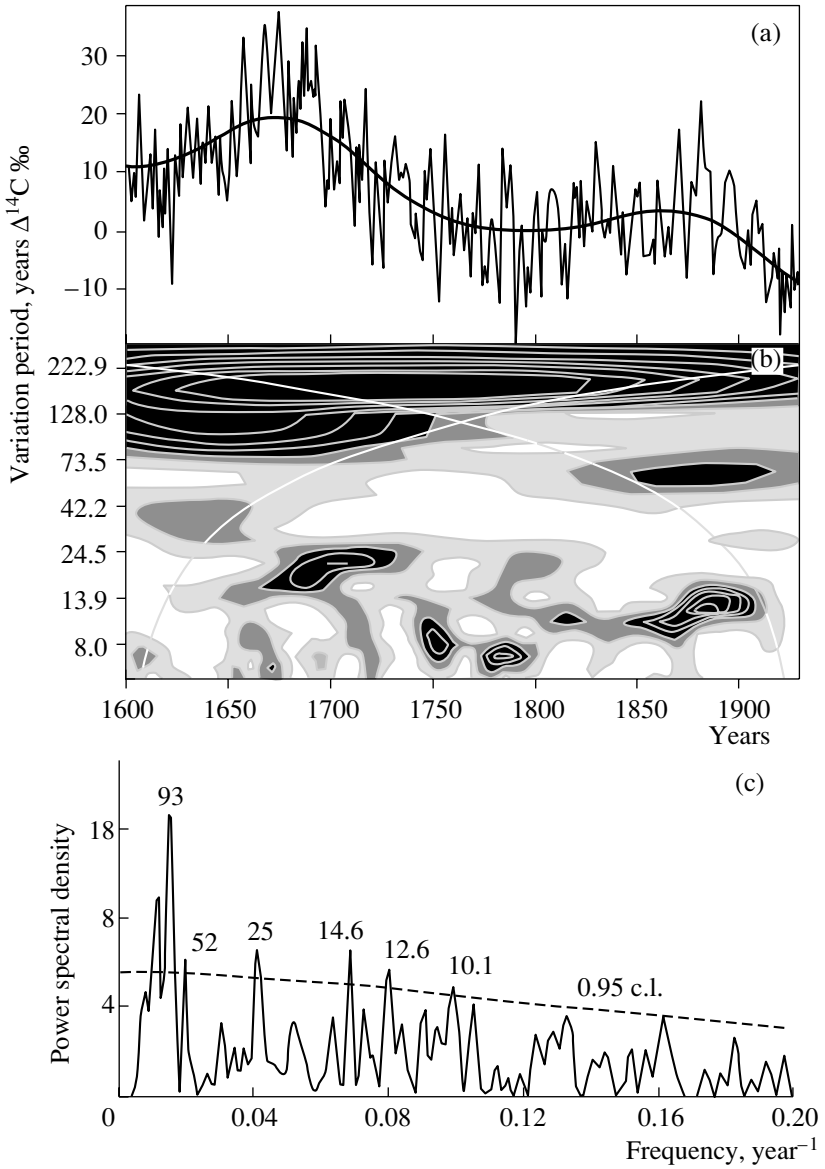

Fig. 1. (a) Radiocarbon concentration measured at TSU (the light curve shows the initial series, and the bold curve the long-timescale trend). (b) Local wavelet spectrum of the TSU series. The grey and black areas show regions above the 0.99 confidence level. (c) Normalized Fourier spectrum of the TSU series with the long-timescale trend subtracted. The dashed curve shows the 0.95 confidence level (for red noise with $\alpha = 0.3$).

This paper is dedicated to a statistical analysis of the Tbilisi time series using the most modern mathematical methods.

Figures 1a and 1b present the TSU radiocarbon series and its local wavelet spectrum (using the Morlet wavelet). The significance of the features of the wavelet spectrum was estimated using the method described in [6]. The brightest feature in the Tbilisi series is a powerful quasi-two-hundred-year variation. The spectral–temporal characteristics of the high-frequency (< 30 years) part of the ^{14}C spectrum can be readily traced in Fig. 1b. We can see from this figure that

(1) a significant 10–14-year periodicity has dominated since the beginning of the 19th century,

(2) there are variations with a period of 6–10 years in the second half of the 18th century,

(3) during the Maunder minimum, there is a variation of $\Delta^{14}C$ with a period of 15–24 years.

Thus, the results of the wavelet analysis of the TSU data are in good agreement with the results of the spectral–temporal analysis carried out by Kocharov and Peristykh [7] using the moving periodogram method. There are some discrepancies only during the Maunder minimum: the local wavelet

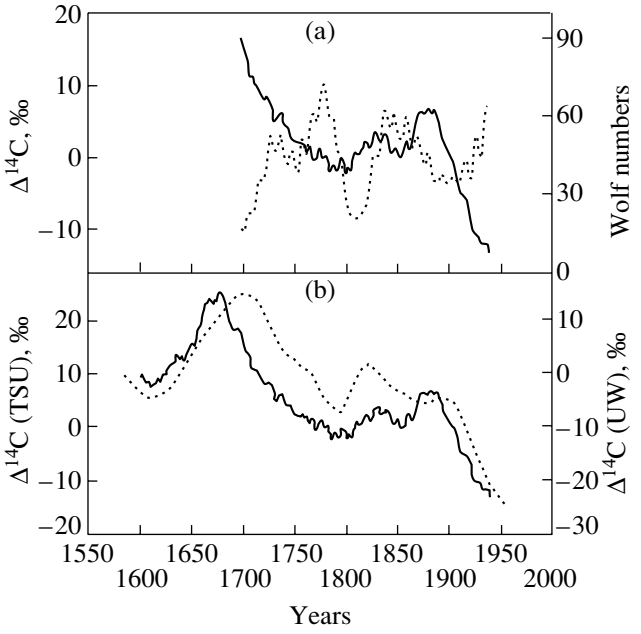

Fig. 2. (a) Wolf numbers (dashed) and TSU radiocarbon concentration (solid), both smoothed over 25 points; (b) radiocarbon concentration measured at TSU (dashed) and UW (solid), both smoothed over 25 points.

spectrum does not show the appreciable peaks in the decade range visible in the spectral–temporal analysis diagrams of [7]. The Fourier spectrum of the high-frequency part of the Tbilisi series obtained by subtracting the long-timescale trend from the initial series (bold curve in Fig. 1a) is presented in Fig. 1c. As we can see, the average power spectral density over the entire interval 1600–1940 also displays pronounced peaks with periods near 25 years and 10–15 years.

Thus, our statistical analysis of the TSU series suggests a rather efficient SA-associated modulation of the radiocarbon concentration. It is quite natural to associate the 200-year variation of Δ^{14}C with the Suess quasi-two-hundred-year solar cycle, the 16–25-year periodicity with the quasi-twenty-two-year Hale cycle, and the 10–15-year variations with the Schwabe quasi-eleven-year cycle. To clarify this question, we compared the TSU series with the Wolf-number record (the longest data series related to solar activity) and the radiocarbon series of [8], obtained at the University of Washington (UW; Seattle, USA) using trees that grew in the northwest of the United States ($\approx 47°$ N).

Figure 2a presents the Wolf numbers and TSU data smoothed over 25 years. The correlation coefficient between these smoothed series reaches -0.33 for the period 1700–1990. The significance of the correlation, estimated using the statistical analysis described in the Appendix, is 0.89. The correlation coefficient for the Tbilisi and Washington series smoothed over 25 years, presented in Fig. 2b, is 0.72 (the significance is higher than 0.99). It is obvious that the TSU Δ^{14}C series is fairly well anticorrelated with the solar activity and well correlated with the ^{14}C concentration in the UW series on timescales of several decades and longer. This confirms the existence of a clear connection between long-timescale changes in the concentration of atmospheric radiocarbon and corresponding variations in solar activity.

The short-timescale, or high frequency ($T <$ 30 years), variations of the ^{14}C concentration in the Tbilisi data display a good negative correlation with the Wolf numbers with a phase shift of five years beginning in the middle of the 19th century (Fig. 3a). In 1845–1930 (7–8 Schwabe cycles), the correlation coefficient is -0.52 (with a significance higher than 0.99). The phase shift of five years is essentially consistent with the time delay between the 11-year variations of the GCR intensity and variations of the ^{14}C concentration in the troposphere (two to four years), estimated theoretically and experimentally in [8–12]. The radiocarbon variations with 6–11-year periods in the 18th century and the first half of the 19th century are weakly correlated with the 11-year solar cycle. The high-frequency variations of Δ^{14}C in the UW series do not show any appreciable correlations with the corresponding radiocarbon variations in the TSU series.

Figure 3b shows that some peaks of the ^{14}C concentration in the UW series are positively correlated with peaks of the TSU series, while some peaks are negatively correlated, with the average correlation for the entire interval 1600–1940 being close to zero. Another feature of the short-timescale variations of the ^{14}C concentration in the TSU measurements is their considerable amplitudes, which substantially exceed the amplitudes of the corresponding Δ^{14}C variations in the UW series (see Fig. 3b). According to [13], the peak-to-peak amplitude of the quasi-twenty-two-year Δ^{14}C variations during the Maunder minimum is close to 9%, and the amplitude of the quasi-eleven-year variations in the 19–20th centuries reaches 2.0–2.4%. Let us estimate the amplitude of the variations of the GCR intensity required to yield such strong variations of the tropospheric radiocarbon concentration. We first determine the amplitude of the required variations of Q, the rate of generation of ^{14}C in the atmosphere. It is known that the carbon-exchange system suppresses high-frequency variations of ^{14}C concentration. The suppression factor for the 11-year variations is approximately 100, while that for the 22-year variations is approximately 50 [9, 10]. This means that the 22-year

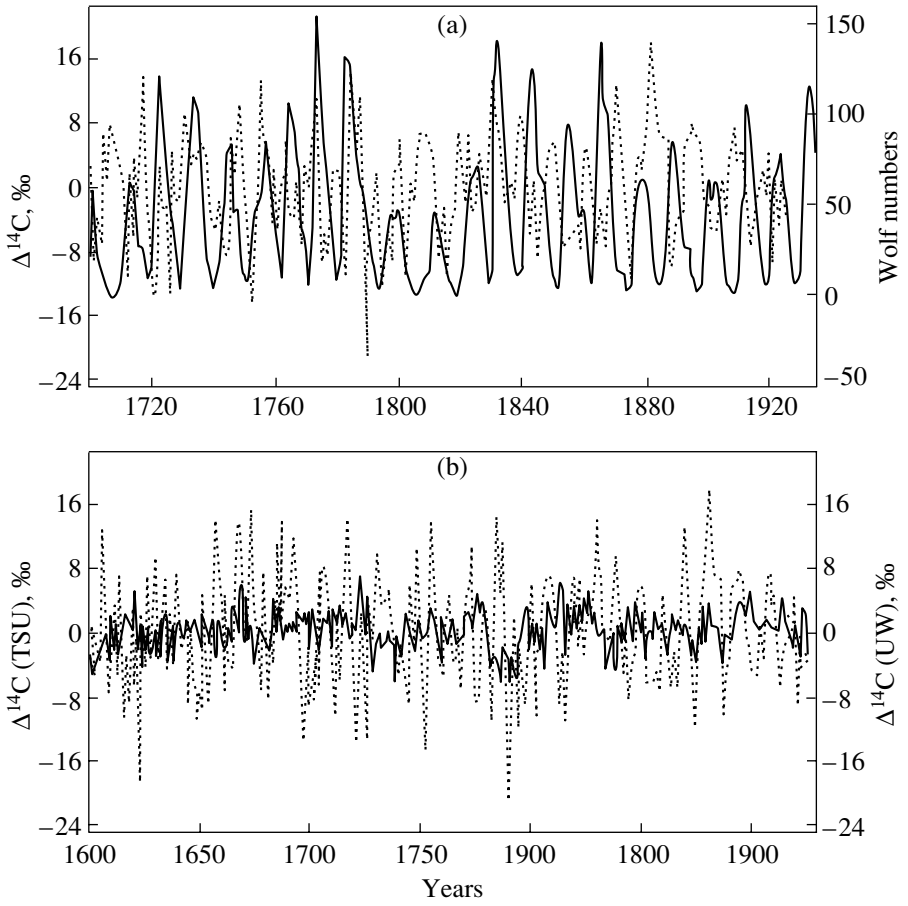

Fig. 3. (a) Short-timescale variations of the radiocarbon concentration measured at TSU (dashed) and the Wolf numbers shifted by five years (solid); (b) short-timescale variations of the radiocarbon concentration measured at TSU (dashed) and UW (solid).

variations of the tropospheric radiocarbon concentration (amplitude $9‰$) corresponds to variations of the carbon generation rate in the atmosphere Q by 45%, while the 11-year variations (amplitude 2.0–2.4%) correspond to a variation of Q by 200–240%. It is now easy to estimate the change in the GCR intensity required to cause such variations of Q. For this purpose we use the formula [14]

$$\Delta I = -4.2 \times \Delta Q^{0.78},$$

where ΔI is in percent of 3854 m^{-2} s^{-1} sr^{-1}, ΔQ is in percent of 2.54 atm cm^{-2} s^{-1} (the average values of I and Q in 1890), and I varies in the energy range 0.5–50 GeV. For $\Delta Q = 45\%$, this formula yields an 80% peak-to-peak variation in ΔI.

Such powerful variations of the GCR intensity could hardly have taken place during the Maunder minimum. Indeed, Hoyt and Schatten [15] have shown that fluctuations of the annual average number of sunspot groups did not exceed two to four during the Maunder minimum. This means that the diffusion mechanism for the solar modulation of the cosmic-ray intensity, which is associated with scattering on magnetic-field irregularities and provides (30–40)% variations of the GCR flux at epochs of normal solar activity, was considerably weakened in 1645–1715.

Another mechanism for the modulation of the GCR intensity in the heliosphere is associated with the drift of particles in the interplanetary magnetic field, which reverses its polarity every 22 years. According to [16], at epochs of deep minima of solar activity, the drift modulation can yield variations of the GCR flux at energies of 0.5–50 GeV with amplitudes of up to 15%. However, this is a factor of five lower than the amplitude required to maintain the 9% variations of Δ^{14}C. A still greater difference between the theoretical and experimental values of the amplitude arises for the 11-year Δ^{14}C cycle in the 19th and 20th centuries. Estimates demonstrate that 11-year variations of the ^{14}C concentration with a peak-to-peak amplitude of 2.0–2.4% should result from approximately 300% variation of the GCR intensity.

It is obvious that the short-timescale variations of the radiocarbon concentration in the TSU series cannot be due to fluctuations in the GCR intensity alone.

Thus, the high-frequency variations of $\Delta^{14}C$ in the Tbilisi series

(1) are not correlated with the high-frequency radiocarbon variations in the University of Washington series,

(2) have very large amplitudes, which cannot be explained by SA-modulated variations of the GCR intensity.

The latter effect is especially pronounced for the quasi-eleven-year cycle. This was discovered in 1983 by Akatova and Kocharov [17], who showed that the very large amplitudes of the 11-year variations of $\Delta^{14}C$ in tree rings and wines cannot be explained by solar modulation alone.

Here, we should note that the character of terrestrial manifestations of the 11-year solar cycle is often poorly predicted by the theory. As one example, we can cite a series of measurements of the concentration of cosmogenic ^{10}Be in South Greenland ice obtained by Beer *et al.* [18], covering the interval 1423–1985. It is known that cosmogenic beryllium, which, like radiocarbon, is produced in the Earth's atmosphere under the action of energetic cosmic rays, is a fairly reliable indicator of SA variations. It was found in [19] that the 11-year periodicity in the ^{10}Be concentration was also preserved during the Maunder minimum, with its amplitude in this period being no smaller than during epochs of normal solar activity. The maximum annual average number of sunspot groups did not exceed two to four in 1645–1715 [15]; i.e., they were approximately a factor of 50 less than in a normal Schwabe cycle. According to [20], the variations of the sunspot group numbers at this time were not cyclic.

Another example is the series of data on the concentration of the stable isotope of oxygen, ^{18}O, in Greenland ice obtained by Stuiver *et al.* [21]. This isotope is known to be a good indicator of the regional climate. Stuiver *et al.* [21] found an 11-year periodicity of ^{18}O in the 18th–20th centuries, which is well correlated with corresponding variations of the Wolf numbers and ^{10}Be. However, the amplitude of the 11-year cycle in $\delta^{18}O$ turned out to be $1.54\,{}^0/_{00}$, whereas theoretical estimations yielded only $0.54\,{}^0/_{00}$, almost three times lower. It is obvious that the quasi-eleven-year cycles demonstrate behavior that cannot be accounted for by solar activity alone, not only in the Tbilisi radiocarbon series, but also in the beryllium and oxygen series. Therefore, some authors have focused on climatic changes as a probable additional source of variations at these frequencies. Usoskin *et al.* [22] have attributed the strong 11-year variations in the beryllium concentration during the Maunder minimum to local climatic fluctuations not related to solar activity.

Stuiver *et al.* [21] have indicated a possible amplification of the 11-year variations of ^{18}O by decade-scale variations in the ocean–atmosphere system and in volcanic activity. Note that variations with periods close to 11 years are quite often revealed in regional climatic parameters. For instance, according to Pudovkin and Morozova [23], 11-year variations of temperature, precipitation, and atmospheric pressure were observed in Switzerland during the 18th–20th centuries. In this connection, it is appropriate to consider whether climatic factors can contribute considerably to short-timescale variations of the radiocarbon concentration in the atmosphere. This problem is also quite topical in connection with the 15–30-year variations of $\Delta^{14}C$, apparently not related to solar activity, that had already been found by Rivin [24] in the Washington radiocarbon series.

Thus, we will consider to what extent such variations could be related to the regional climate. According to theoretical estimates [10, 25], the atmospheric concentration of ^{14}C can be affected by long-timescale (with periods of hundreds of years and longer) global climatic variations. The effect of shorter-timescale climatic changes should be weaker [10]. The manifestations of local climatic effects should be especially weak, since radiocarbon spends only a few years in the atmosphere, and the time for the mixing of the CO_2 of the lower troposphere in each hemisphere is only several months [26]. Nevertheless, definite evidence that the global and regional climate affects the short-timescale variations of the radiocarbon concentration in tree rings was obtained in [13, 27]. In particular, Vasil'ev *et al.* [13] detected a clear quasi-eleven-year cycle in the indices of tree-ring growth for trees from the south of Karelia, the region from which the wood for the second part of the Tbilisi series (1800–1940) was taken. In [13] the climatic factor was removed from the radiocarbon concentration measured in [5] from the Karelian dendrodata. This procedure was based on a simple linear regression, and resulted in a considerable (no less than one-third) decrease of the amplitude of the 11-year variation of $\Delta^{14}C$.

Further evidence for the possible effect of the local climate on the ^{14}C concentration was obtained by Damon *et al.* [28], who found an appreciable correlation between the radiocarbon concentration in the rings of trees that grew in the region of the Mackenzie River and the local average air temperature for May–August. Such evidence should not be neglected, though the idea that there could be a relation between the atmospheric radiocarbon and

regional climate does not completely match generally accepted concepts about the carbon-exchange cycle. If this hypothesis is correct and a significant regional climatic component is indeed present in the high-frequency fluctuations of the tropospheric radiocarbon, then the lack of a correlation between the short-timescale variations of $\Delta^{14}C$ in tree rings from the northwest of the USA and the east of Europe (the Carpathians, Karelia) can be explained in a natural way.

Thus, if we accept the above supposition as a working hypothesis, we can represent the 11-year periodicity in $\Delta^{14}C$ measured in [5] as a sum of oscillations: (a) the 11-year cycle in the rate of generation of atmospheric radiocarbon Q, directly connected with the Schwabe cycle in the GCR intensity, and (b) powerful decade-scale variations of atmospheric ^{14}C associated with corresponding variations of the regional climate. The decade climatic oscillations could, in turn, also be associated with the 11-year solar cycle. The GCR-driven 11-year variations of Q (direct solar modulation) are very clearly phase-locked to the solar cycle. The phase link between the quasi-eleven-year climatic cycle and the Schwabe cycle should not necessarily be constant in time. For instance, it was shown in [23] that the 11-year cycles in temperature, precipitation, and atmospheric pressure are positively correlated with the corresponding variations of the solar activity in some periods, while they are negatively correlated in other periods. Thus, the 11-year periodicity in $\Delta^{14}C$ could be either amplified and undistorted if the cycles in Q and in the local climate stably coincide in phase, or appreciably distorted and even weakened if the cycles in Q and in the local climate have different phases. It is reasonable to suppose that, in the experimental series [5], the former case was realized in 1850–1940 and the latter in 1700–1850.

Recently, evidence has grown suggesting that the GCR intensity is precisely the factor that is physically transferring the effect of the Sun to the Earth's climate [29–30]. If this is the case, the ^{14}C variations in the troposphere could be related to changes in solar activity and the cosmic-ray flux through two chains.

(1) Variations of solar activity ⟶ variations of the GCR intensity ⟶ variations of the ^{14}C production rate in the atmosphere ⟶ variations of the ^{14}C concentration in the troposphere.

(2) Variations of solar activity ⟶ variations of the GCR intensity ⟶ variations of the global and regional climate ⟶ variations of the ^{14}C concentration in the troposphere.

Of course, in this case, too, short-timescale fluctuations of the tropospheric radiocarbon contain information on the corresponding variations of the GCR flux; however, this information is not direct, and is mediated to a considerable extent by poorly understood climatic factors. Such information could be suitable for qualitative conclusions—for example, about the presence in a signal in a specified interval of variations with a definite period—but is not suitable for quantitative analyses (the amplitude of variations, phase coupling of the signal with other data series, etc.). In the hypothesis under consideration, the considerable increase in the period of high-frequency variations of $\Delta^{14}C$ during the Maunder minimum (up to 15–24 years) can be considered evidence for 22-year variations of the GCR intensity in this interval, though any quantitative estimates of this effect based on the data of [5] are unlikely to be very reliable. However, even the qualitative conclusion that there was a 22-year modulation in the GCR intensity during the Maunder epoch is extremely important.

Since the Hale solar magnetic cycle is usually associated with the activity of the dynamo mechanism, its preservation in 1645–1715 indicates that the solar dynamo continued to operate at that time, and that solar convection did not cease. This demonstrates the fundamental character of the solar dynamo. In this case, it is reasonable to explain the deep Maunder minimum of solar activity as a superposition of the minima of several long-timescale cycles [31]. However, we must also note that available information about the time behavior of the solar activity during the Maunder minimum is quite contradictory. Radiocarbon measurements carried out at the Ioffe Physicotechnical Institute [14] confirm the presence of 22-year variations of the GCR intensity during the Maunder minimum.

The interpretation of the UW experimental series is not unambiguous. In their analysis of these data, Stuiver and Braziunas [8] noted a possible 11-year cycle during the Maunder minimum, while Peristykh and Damon [32] concluded that 22-year variations dominated in 1630–1720, and Ogurtsov and Vasil'ev [27, 33] found a 16-year periodicity in the UW series for 1645–1715. Recall that the concentration of cosmogenic beryllium displays strong variations with a period close to 11 years during the Maunder minimum [18]. Note that analyses of solar cyclicity during the Maunder minimum based on cosmogenic isotopes are also complicated by the fact that the Cassiopeia A supernova explosion, which probably occurred in about 1672, could provide an additional enhancement in the concentrations of these isotopes, distorting the solar–terrestrial variations [34]. However, apart from cosmogenic isotopes, quasi-twenty-year variations during the Maunder minimum have also been derived from the concentration of nitrates

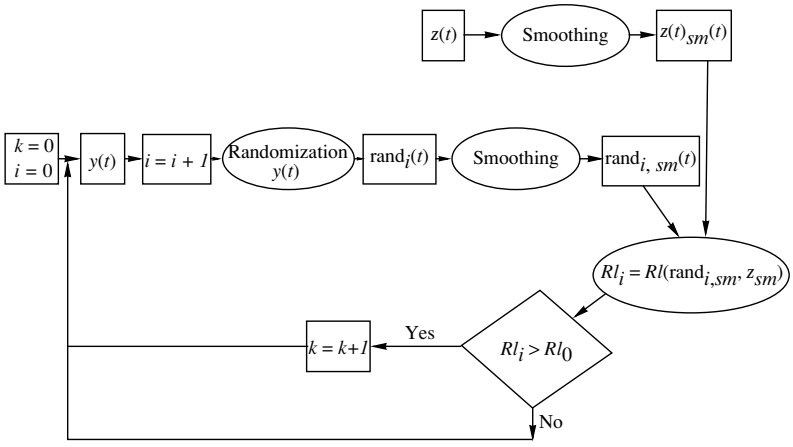

Fig. 4. Scheme of the statistical experiment used to estimate the significance of a correlation between smoothed series of data.

(NO_3^- ions) in Greenland ice [35, 36], which Gladysheva *et al.* [36] have attributed completely to the corresponding GCR cycle.

Summarizing our study of the Tbilisi radiocarbon series, we can draw the following conclusions.

(1) Long-timescale (with periods of tens and hundreds of years) variations of the radiocarbon concentration in tree rings reflect the corresponding variations of the GCR intensity. This makes atmospheric radiocarbon a reliable indicator of global minima and maxima of solar activity.

(2) Short-timescale variations of the ^{14}C concentration depend on variations of the GCR intensity and solar activity in a more complex manner. We have evidence for the presence of a significant component in the high-frequency part of the radiocarbon spectrum that is not associated directly with solar activity, and is probably related to variations of the regional climate. These variations distort the radiocarbon signal, making it suitable only for qualitative analyses of short-timescale variations of the solar activity and cosmic-ray intensity.

(3) The Tbilisi radiocarbon series contains indications of 22-year variations of the GCR flux during the Maunder minimum of the solar activity. In turn, this testifies that the polarity of the interplanetary magnetic field continued to reverse every 22 years in 1645–1715; i.e., the dynamo mechanism was not completely disabled during the epoch of the Maunder minimum. In this case, the prolonged and deep SA minimum in 1645–1715 could be logically understood as being due to the coincidence of the minima of several long-timescale solar cycles. However, it is premature to make any definite conclusions on the behavior of the solar activity during the Maunder minimum, as the information from different sources is rather contradictory.

It is clear that analyses of solar activity in the past based on the concentrations of cosmogenic isotopes in natural "archives" are a reasonably promising direction in paleoastrophysics, although this problem is probably more difficult than was supposed earlier. The joint analysis of several data series on the concentrations of cosmogenic isotopes with different formation mechanisms could potentially be quite fruitful—an approach that was initiated in [37]. In any case, the elucidation of the roles of global and local climatic factors in determining the concentrations of cosmogenic isotopes is clearly an important direction for future work in this field.

ACKNOWLEDGMENTS

This work was carried out in the framework of the program of scientific exchange between the Russian and Finnish Academies of Sciences (project No. 16) and the program of the Presidium of the Russian Academy of Sciences "Non-stationary Phenomena in Astronomy." This work was supported by INTAS (grant 2001-0550) and the Russian Foundation for Basic Research (project nos. 03-04-48769 and 03-02-17505).

APPENDIX

ESTIMATION OF THE SIGNIFICANCE OF THE CORRELATION BETWEEN SMOOTHED DATA SERIES

If two smoothed data series $y(t)$ and $z(t)$ are found to be correlated with some correlation coefficient R_0, there remains the problem of estimating the significance of this correlation. We solved this problem by performing a statistical experiment in which we adopted as the null hypothesis the hypothesis that one of the series ($y(t)$ or $z(t)$) is noise (white or red)

and that the observed correlation arose by chance. We then tested this null hypothesis via Monte-Carlo simulations using the scheme shown in Fig. 4.

We first randomize (mix randomly) one of the series ($z(t)$ in Fig. 4) and construct from it a noise series $\mathrm{rand}_i(t)$, where $y(t)$ is considered at this stage to be the useful signal. The noise is constructed according to the well-known relation

$$\mathrm{rand}_i(t) = \alpha \times \mathrm{rand}_{i-1}(t) + \epsilon(t),$$

where $\epsilon(t)$ is the random version of the initial signal obtained by shuffling it and α is the red-noise index. The main advantage of using such noise is that it has a distribution function that either precisely coincides with that of the analyzed series ($\alpha = 0$), or is at least quite similar to it ($\alpha > 0$). This is important, because the noise in real experiments can be substantilly non-Gaussian, as, e.g., in the Davis chlorine–argon neutrino experiment. We smooth the noise $\mathrm{rand}_i(t)$ obtained in this way and calculate the correlation coefficient Rl_i between the smoothed series $\mathrm{rand}_i(t)$ and $y(t)$. If Rl_i is greater than R_0, the result of the trial is considered to be positive. After realizing N trials, the significance of the correlation coefficient R_0 is determined as $1.0 - K^+/N$, where K^+ is the number of trials with a positive result. This procedure is then repeated with $z(t)$ as the useful signal and constructing the noise from $y(t)$. As a result, we obtain two estimates of the significance, and choose the smaller as the final estimate. It is obvious that, if we treat both data series as noise and randomize them both, the confidence level will be higher. Thus, this type of statistical experiment yields a lower limit for the significance level of the correlation.

REFERENCES

1. G. E. Kocharov, Nucl. Phys. B (Proc. Suppl.) **22B**, 153 (1991).
2. V. A. Dergachev, Izv. Ross. Akad. Nauk, Ser. Fiz. **59** (7), 53 (1995).
3. M. Stuiver, T. F. Braziunas, B. Becker, and B. Kromer, Quat. Res. **35**, 124 (1991).
4. B. P. Konstantinov and G. E. Kocharov, Dokl. Akad. Nauk SSSR **165**, 63 (1965).
5. I. V. Zhorzholiani, P. G. Kereselidze, G. E. Kocharov, *et al.*, *Experimental Methods for Investigating Astrophysical and Geophysical Phenomena* (FTI, Leningrad, 1990) [in Russian], p. 92.
6. M. G. Ogurtsov, Yu. A. Nagovitsyn, G. E. Kocharov, and H. Jungner, Solar Phys. (2003, in press).
7. G. E. Kocharov and A. N. Peristykh, *Possible Methods for Measuring Very Small Amounts of Isotopes* (FTI, Leningrad, 1990) [in Russian], p. 5.
8. M. Stuiver and T. Braziunas, Holocene **3**, 289 (1993).
9. G. E. Kocharov, *Astrophysical Phenomena and the Concentration of Radiocarbon in Terrestrial Samples* (St. Petersburg, 1996) [in Russian].
10. V. A. Dergachev and V. S. Veksler, *Application of Radiocarbon Methods for Studies of the Environment in the Past* (Leningrad, 1991) [in Russian].
11. P. E. Damon, A. Long, and E. I. Wallick, Earth Planet. Sci. Lett. **20**, 300 (1973).
12. P. Povinec, Acta Phys. Connen. **18**, 139 (1977).
13. S. S. Vasil'ev, G. E. Kocharov, and M. G. Ogurtsov, Izv. Ross. Akad. Nauk, Ser. Fiz. **61**, 1224 (1997).
14. V. A. Vasil'ev and G. E. Kocharov, in *Proceedings of XIII Leningrad Seminar on Cosmophysics, Leningrad, 1983*, p. 75.
15. D. V. Hoyt and K. Schatten, Sol. Phys. **179**, 189 (1998).
16. G. E. Kocharov, V. M. Ostryakov, A. N. Peristykh, and V. A. Vasiliev, Sol. Phys. **159**, 381 (1995).
17. N. I. Akatova and G. E. Kocharov, Preprint No. 802 (Phys. Tech. Inst., Russ. Acad. Sci., 1983).
18. J. Beer, St. Baumgartner, B. Dittrich-Hannen, *et al.*, in *The Sun as a Variable Star: Solar and Stellar Irradiance Variations,* Ed. by J. M. Pap, C. Frohlich, H. S. Hudson, and S. K. Solanki (Cambridge Univ. Press, Cambridge, 1994), p. 291.
19. J. Beer, S. Tobias, and N. Weiss, Sol. Phys. **181**, 237 (1998).
20. P. Frick, D. Galyadin, D. Hoyt, *et al.*, Astron. Astrophys. **338**, 670 (1997).
21. M. Stuiver, P. Grootes, and T. Braziunas, Quat. Res. **44**, 341 (1995).
22. I. G. Usoskin, K. Mursula, and G. A. Kovaltsov, J. Geophys. Res. **106** (A8), 16039 (2001).
23. M. I. Pudovkin and A. L. Morozova, Geomagn. Aeron. **40** (3), 3 (2000).
24. Yu. A. Rivin, Geomagn. Aeron. **38**, 183 (1998).
25. P. E. Damon, J. C. Lerman, and A. Long, Ann. Rev. Earth Planet. Sci. **6**, 457 (1978).
26. G. I. Pearman and P. Hyson, J. Geophys. Res. **85**, 4457 (1980).
27. M. G. Ogurtsov and S. S. Vasiliev, in *Views and News in Physics and Astrophysics* (Helsinki–St. Petersburg, 1997), p. 82.
28. P. E. Damon, G. Burr, A. N. Peristykh, *et al.*, Radiocarbon **38** (3), 324 (1996).
29. N. Marsh and H. Svensmark, Phys. Rev. Lett. **85** (23), 5004 (2000).
30. O. M. Raspopov, O. I. Shumilov, and E. A. Kasatkina, Biofizika **43** (5), 902 (1998).
31. G. E. Kocharov, Nuovo Cimento **19**, 883 (1996).
32. A. N. Peristykh and P. E. Damon, Sol. Phys. **177**, 343 (1998).
33. S. S. Vasiliev and M. G. Ogurtsov, Preprint No. 1695, PhTI RAN (Ioffe Phys.-Techn. Inst., Russ. Acad. Sci., 1997).

34. G. E. Kocharov, Yu. N. Markov, R. Ya. Metskhvarishvili, and S. L. Tsereteli, in *Proceedings of XIII Leningrad Seminar on Cosmophysics, Leningrad, 1983*, p. 102.
35. M. G. Ogurtsov, Candidate's Dissertation in Physics and Mathematics (St. Peterburg, 1997).
36. O. G. Gladysheva, G. E. Kocharov, G. E. Kovaltsov, and I. G. Usoskin, Adv. Space Res. **29** (11), 1707 (2002).
37. Yu. A. Nagovitsyn and M. G. Ogurtsov, Izv. GAO **215**, 271 (2000).

Translated by G. Rudnitskiĭ

Astronomy Reports, Vol. 47, No. 12, 2003, pp. 975–978. Translated from Astronomicheskiĭ Zhurnal, Vol. 80, No. 12, 2003, pp. 1059–1063.
Original Russian Text Copyright © 2003 by Sokoloff.

Topology of the Universe and Topological Inflation

D. D. Sokoloff

Moscow State University, Vorob'evy gory, Moscow, 119899 Russia

Received May 14, 2002; in final form, May 8, 2003

Abstract—A cosmological scenario in which the topology of the Universe is treated like a dynamical time-dependent variable is put forward. The Universe could be small in an initial quantum stage of evolution and then gradually increase its dimensions so that the present-day nontrivial topology is manifest only far beyond the cosmological horizon. © *2003 MAIK "Nauka/Interperiodica"*.

1. INTRODUCTION

Despite the obvious successes of modern cosmology, it still involves certain unclear basic elements arising in the logical structure of the theory of general relativity itself. This concerns the principles of predicting the topology of the Universe. The standard cosmological approach is based on deriving the metric tensor from certain differential equations, namely, the Einstein equations. To write these equations, we must assume that the corresponding independent variables are specified in some way. In other words, the topological structure of the manifold supporting the required metric is specified. Precisely this structure enables us to assign meaning to the differential operations appearing in the Einstein equations, and accordingly to pose the problem of how they are realized; i.e., the problem of finding the corresponding metric. From this point of view, the choice of a topology for a cosmological model should be based on some principles lying outside general relativity.

In practice, cosmology is not a logically consistent theory, and the construction of cosmological models begins with a local analysis of the Einstein equations within a single map that covers only some part of the cosmological model under consideration. The construction of a complete cosmological model is based on some procedure for continuing a solution of the Einstein equations, similar to the continuation of an analytic function (see, for example, [1, 2]). Simple examples show that the formalization of this procedure is unlikely to be successful. In fact, even a small piece of a Euclidean plane can be continued to either a complete plane, a cone, or a polyhedral metric without vertices. It is difficult to justify the choice of a plane from the point of view of completeness, since the desired cosmological models can obviously be incomplete (this is precisely the phenomenon of the Big Bang). The well-known example of the continuability of the Taub–NUT metric [3] to an object violating the Hausdorff separation axiom demonstrates that much more exotic situations are also possible.

Practical cosmology avoids such topological problems by considering only Friedmann-like models of the Universe. These are based on an initial separation of space–time into a space-like cross section and a time axis, with the spatial section being uniform and isotropic. As a result, the determination of the scale factor (i.e., the time evolution) and of the topological structure of the Universe will be relatively isolated problems. The set of admissible topologies becomes quite limited, and choosing a realistic model can be treated as an observational problem. This is precisely how cosmology has developed since the time when the problem of choosing the topological structure of cosmological models was understood (in fact, just after the formulation of Einstein's approach to cosmology, based on general relativity).

As is known, no appreciable deviations from the simplest Euclidean topology are manifest in virtually the entire observable part of the Universe. The first observational estimates of the region of trivial topology were obtained in the 1970s (for example, by Sokolov and Shvartsman [4]). It was found that the available observational data confirmed the trivial topology only within a relatively small (from the cosmological point of view) distance. This situation changed substantially in the 1990s, with the discovery of fluctuations of the cosmic microwave background radiation. As a result, we can now be sure that the Universe possesses a simple topological structure almost out to the horizon (see, e.g., the review by Lachiéze-Rey and Luminet [5]).

Although this progress is remarkable, it does not enable us to explain *why* the Universe has such a topology. Moreover, it is interesting to construct a compact, spatially-flat Universe using a nontrivial topology [6].

1063-7729/03/4712-0975$24.00 © 2003 MAIK "Nauka/Interperiodica"

It is likely that a self-consistent solution of the above problem can be obtained only by going beyond the framework of general relativity. We have not accumulated a sufficient amount of observational information to carry out such a generalization. However, the internal difficulties of formulating a procedure for observationally determining the topological structure of the Universe are beginning to be discussed in the literature [7, 8]. Therefore, to some degree, discussions of various theoretical approaches to revealing the topology of the Universe are beginning to make the transition from pure fantasy to the realm of reasonable scientific problems. If we wish to stray from the traditional principles of general relativity as little as possible, the admissible modifications are fairly restricted, and a sketch of the future theory can be outlined. This is the aim of the present paper.

2. THE VARIATIONAL PRINCIPLE AND THE TOPOLOGY OF THE UNIVERSE

We should first point out that the variational formulation of general relativity appears to be most favorable for attacking the problem of the topology of the Universe. In this formulation (see, for example, [9, p. 343]), a gravitational field is determined by the stationary condition of the action functional

$$S_g = \int L\sqrt{-g}d\Omega, \qquad (1)$$

where L is the Lagrangian, $d\Omega$ is a volume element, g is the determinant of the metric tensor, and the integration is carried out over the entire spatial section and over the time coordinate between two specified values. Note that space–time is implicitly assumed here to be topologically factorizable into space and time. The quantity S_g is commonly treated as a functional of the metric tensor g_{ij}, so that its variation with respect to the coordinates yields the Einstein equations. Of course, the functional S_g also depends on the choice of topological structure of the spatial cross section. Therefore, if it is possible to vary S_g with respect to quantities specifying the spatial section, it may be possible to obtain equations whose solution determines the topology of the Universe. This is the basic idea of our method.

The main obstacle is that the set of all possible topologies of the cosmological model does not represent a manifold, i.e., it cannot be parameterized by several continuous parameters, and it is not clear how the functional should be varied. We shall choose the simplest way to resolve this difficulty, and assume that the cosmological model under consideration is spatially flat and possesses the topology of a three-dimensional torus. In addition, we will assume the directions of the topological gluing to be fixed. Therefore, we shall consider cosmological models of the form

$$ds^2 = c^2dt^2 + a^2(t)(dx^2 + dy^2 + dz^2), \qquad (2)$$

$$x = x + A, \quad y = y + B, \quad z = z + C. \qquad (3)$$

This action functional depends on the scale factor a and three scalar quantities A, B, and C. Variation with respect to a results in the ordinary equation following from the Einstein equations, while variation with respect to A, B, and C can lead to three dynamic equations that determine the topology.[1] However, the gravitational-field Lagrangian does not depend on the topological variables A, B, and C, so that the functional S_g is proportional to ABC and does not possess an extremal, so that the resulting equations are incompatible. This probably means that it is impossible to determine the topology of the Universe within the framework of classical general relativity; however, it may be possible in quantum cosmology.

3. THE CASIMIR EFFECT AND EVOLUTION OF THE TOPOLOGY

Quantum effects violate the lack of dependence of the Lagrangian L on the topological variables A, B, and C. The best-known phenomenon of this kind is the Casimir effect. Its essence is that a Universe satisfying the gluing conditions (3) can be treated like a huge capacitor, whose field Lagrangian is slightly different from the Lagrangian in a vacuum. It is obvious that these deviations are appreciable only in the earliest stages of evolution of the Universe, when its radius of curvature is comparable to the Planck scale. Under such conditions, the gravitational-field Lagrangian acquires a quantum correction [6], whose terms depend on the topological variables and are proportional to the Planck constant $\hbar$ (or to some power of this constant):

$$L = L_E(a) + \hbar L_q(a, A, B, C). \qquad (4)$$

As a result, the extremum condition can lead to nontrivial algebraic equations of the form $F(A, B, C) = 0$ for the topological parameters. Precisely their solution will specify the true values $A = A^*$, $B = B^*$, and $C = C^*$.

These ideas have not been substantially developed over the last two decades, and the analysis of observational data likewise does not suggest any preferred topology. Therefore, let us assume that there are no

[1] From the viewpoint of topology as a mathematical science, all models of this family have the same topological structure but different global structures. This discrepancy between the mathematical and physical terminology is explained in more detail, for example, by Sokolov and Shvartsman [4].

preferred values of the topological parameters ($F \neq 0$). Moreover, physical laws are usually described by dynamic (evolutionary) equations rather than algebraic relations.

To obtain differential equations for the topological parameters by varying the action, L_q must contain $\dot{A}$, $\dot{B}$, and $\dot{C}$, where a dot denotes a time derivative. Previous calculations of L_q were based on the assumption that the topology of the Universe was constant, and so they did not contain time derivatives of A, B, and C. On the other hand, there is no reason why it should not be possible to introduce terms with such derivatives into the theory. Bearing in mind the fact that the corresponding theory has not been fully developed, we will simply suppose that, one way or another, terms containing time derivatives of the topological parameters appear in L_q:

$$L = L_E(a) + \hbar L_q(a, A, B, C, \dot{A}, \dot{B}, \dot{C}, \ldots). \quad (5)$$

As a result, the equations governing the topological variables will take the dynamic form

$$\hbar \ddot{A} F_d + F = 0, \quad (6)$$

where F_d is a term arising due to the variation with respect to the time derivatives.

Since we assumed that $F \neq 0$, the solution of equation (6) varies with a rate inversely proportional to $\hbar^{1/2}$. Since the present size of the Universe is not microscopic, this variation should represent a very rapid increase, so that the initial small size can substantially exceed the size of the modern horizon by the end of the quantum epoch.

Note that Eq. (6) retains its meaning if there is a preferred topology of the Universe. In that case, a solution with a different initial topology will rapidly reach the preferred configuration.

Thus, the variational approach to general relativity provides hints that it is possible to generalize it such that the size of the initially small Universe will sharply grow by the end of the quantum stage of its evolution. As a result, the spatial size of the Universe becomes very large, and exceeds the present-day horizon. In essence, we do not need to consider a quantum–cosmology process that directly gives rise to an infinite Universe: a small Universe can initially be born, after which its dimensions increase to an arbitrarily large value. This process is similar to but appreciably different from inflation. During inflation, the scale factor increases, but all spatial points of the comoving coordinate system (which can be identified, for example, by test particles) exist at all times in the evolution of the Universe. In our scenario, the spacelike cross section of the Universe gradually acquires new regions, whose test particles did not participate in the prior evolution and were created at later times. This process can be called topological inflation.

We emphasize that the rate of increase of the comoving space in the quantum stage of topological inflation could exceed limits imposed by causality, so that the parameters A, B, and C could initially be smaller than the horizon, and then exceed it later.

4. TOPOLOGICAL INFLATION

Of course, the proposed picture represents a quasi-classical approach to quantum cosmology, and we will not attempt to construct a self-consistent theory. We will limit our analysis to elucidating how we can imagine a topological increase of the comoving space.

Let the topological parameter A increase by dA during the time interval dt. As a result, some new spatial points should appear in the cosmological model, which must somehow be assigned new spatial coordinates. It is not possible to simply postulate that these coordinates cover the interval between $x = A$ and $x = A + dA$ because (i) this would violate the spatial uniformity of the x coordinate, and (ii) the coordinate $x = A$ is not distinguished in any way on a flat (zero-curvature) torus. The solution of this conceptual problem may be based on considering a quantum ensemble of universes, whose individual members increase their spatial cross sections at different values of x (as well as of y and z).

Let us describe the corresponding mathematical construction using rather unusual notation. At time $t = t_0$, let the growth in the spatial cross section in x occur near the point $x = \mathfrak{B}_x$, the growth in y occur near $y = \mathfrak{B}_y$, and the growth in z occur near $z = \mathfrak{B}_z$. The wave function of the Universe ♣ will obviously depend on these parameters: ♣ = ♣$(A, \mathfrak{B}_x, B, \mathfrak{B}_y, C, \mathfrak{B}_z)$ (we shall not write explicitly its dependence on other variables). The uniformity of the cosmological model, for example, with respect to x, now reduces to the condition

$$\frac{\partial \clubsuit \clubsuit^*}{\partial \mathfrak{B}_x} = 0, \quad (7)$$

where $(^*)$ represents complex conjugation. In other words, the condition of uniformity implies that the spatial growth occurs in any place with equal probability.

In the proposed scenario, the present state of the Universe is a flat torus, whose three topological parameters can considerably exceed the size of the horizon. These topological parameters are now almost constant, so that the size of the spatial cross section increases only due to variations in the scale factor. Just as a macroscopic body can exhibit with vanishingly small probability traces of quantum motion,

so the modern Universe (in our scenario) can with vanishing small probability form a region of newly born space.

5. DISCUSSION

We have described a scenario that, in principle, can help us resolve one of the conceptual problems of general relativity and cosmology: uncertainty in the choice of the topology of the space-like cross section. The proposed scenario explains why we do not see a nontrivial topological structure of the Universe.

Due to the scarcity of observational data that are potentially able to support our ideas, we have not discussed a particular form of the Lagrangian of this theory. In principle, it is not difficult to write this Lagrangian in a form that is consistent with the constraints available. A substantial limitation of our scenario is that, in general, the topology of the Universe is not described by continuous parameters, so that the action cannot be varied directly with respect to the topology of the space. One way to overcome this obstacle may be to vary the action with respect to the symmetry transformations, whose factorization could lead to a nontrivial topology. The set of corresponding transformations possesses much more regular structure than the set of various topologies for Friedmann cosmological models.

Another advantage of our scenario is that it is not necessary to choose between a universe that is born once or permanently exists—it becomes possible to consider a universe that is continually being born bit by bit. Of course, the next natural step in the development of such a theory is to admit the possibility of the disappearance of some regions of space.

A principal postulate in the construction of our scenario was the assumption that it was possible to factorize (at least, topologically) the Universe into a time axis and a space-like cross section. This seems to be a substantial violation of the basic principles of general relativity. However, simple examples (see, e.g., [10]) show that attempts to preserve the equal status of the temporal and spatial variables even at the topological level open possibilities that are currently far beyond the boldest dreams.

In particular, the variational principle in the proposed scenario is more fundamental than the Einstein equations. The same ideas were put forward by Gertsenshtein and Konstantinov [11], who drew attention to the fact that the action does not diverge in the Friedmann singularity, and used this feature to continue their cosmological model beyond the singularity.

We emphasize that, even if the topological parameters are now considerably larger than the distance to the horizon, they can still affect local physical processes. The presence of topological gluing preserves the local uniformity of Friedmann models, but violates their global isotropy (an exception is a projective space, as a special topological version of a closed Friedmann model [12]). In principle, breaking the global isotropy could lead to the formation of a uniform magnetic field in the quantum stage, when the size of the Universe was microscopic. The corresponding magnetic field lines will be closed but cover a contour that cannot contract to a point. We can imagine that this field will grow together with the comoving space, and this process may help us to resolve another problem of modern cosmology. Our scenario admits the possible existence of a perfectly uniform cosmological magnetic field, but does not indicate a particular path for its formation. On the other hand, even a very weak cosmological magnetic field can substantially modify our understanding of the origin of the seed fields that are amplified by the galactic dynamo mechanism to form the large-scale magnetic fields of spiral galaxies observed today [13].

ACKNOWLEDGMENTS

This work was supported by the Russian Foundation for Basic Research (project no. 01-02-16158). We are grateful to V.I. Noskov for fruitful discussions.

REFERENCES

1. A. Smith, Bull. Austral. Math. Soc. **78**, 141 (1978).
2. V. A. Popov, Mat. Zametki **36**, 559 (1984).
3. C. W. Misner, J. Math. Phys. **4**, 924 (1964).
4. D. D. Sokoloff and V. F. Shvartsmann, Zh. Éksp. Teor. Fiz. **66**, 412 (1974) [Sov. Phys. JETP **39**, 196 (1974)].
5. M. Lachiéze-Rey and J.-P. Luminet, Phys. Rep. **254**, 135 (1995).
6. Ya. B. Zel'dovich and A. A. Starobinskiĭ, Pis'ma Astron. Zh. **10** (1984) [Sov. Astron. Lett. **10**, 135 (1984)].
7. R. Lehouc, J.-P. Uzan, and J.-P. Luminet, Astron. Astrophys. **363**, 1 (2000).
8. D. D. Sokolov, Astron. Zh. **79**, 579 (2002) [Astron. Rep. **46**, 521 (2002)].
9. L. D. Landau and E. M. Lifshits, *Field Theory* (Nauka, Moscow, 1967) [in Russian].
10. M. Firz and R. Jost, Helv. Phys. Acta **38**, 778 (1965).
11. M. E. Gertsenshteĭn and Yu. N. Konstantinov, in *Problems of Gravitation and Elementary Particle Theory* (Atomizdat, Moscow, 1975) [in Russian], Vol. 6, p. 23.
12. D. D. Sokolov, in *Gravitation and Relativity*, **12**, 142 (1977).
13. R. Beck, A. Brandenburg, D. Moss, *et al.*, Ann. Rev. Astron. Astrophys. **34**, 155 (1996).

Translated by Yu. Dumin

Astronomy Reports, Vol. 47, No. 12, 2003, pp. 979–988. Translated from Astronomicheskiĭ Zhurnal, Vol. 80, No. 12, 2003, pp. 1064–1074.
Original Russian Text Copyright © 2003 by Vasiliev, Shchekinov.

The Formation of the First Objects During the Virialization of the Dark-Matter Halos

E. O. Vasiliev[1] and Yu. A. Shchekinov[1,2]

[1] *Physics Department, Rostov State University, Rostov-on-Don, Russia*

[2] *Rostov Branch of Newton Institute, Chili*

Received September 30, 2002; in final form, July 26, 2003

Abstract—The chemical and thermal evolution of the baryon component in the gravitational field of low-mass primordial dark-matter halos during their virialization is studied. We consider low-mass halos to be those for which the characteristic baryon cooling time can appreciably exceed the comoving Hubble time, so that the cooling process can continue to the current epoch ($z \sim 0$). The virialization process is described in two scenarios: "quiet " virialization, in which the establishment of the virial state is assumed to be homogeneous over the entire volume considered, and "violent" virialization, in which the establishment of the virial state is assumed to be realized via the action of shock waves. In this second case, the efficiency of the formation of molecular hydrogen grows substantially, and can reach $H_2/H \sim 0.01$ in some cases, which exceeds current estimates by at least an order of magnitude. This eases the condition for the birth of the first gravitationally bound objects with comparatively low masses ($M \gtrsim 2 \times 10^5 M_\odot$), possibly leading to an appreciable increase in the fraction of the mass contained in Population III objects, and also to a shift in the onset of the formation of the first stars toward higher redshifts. © *2003 MAIK "Nauka/Interperiodica"*.

1. INTRODUCTION

One of the key questions in modern cosmology is the nature of the first stellar objects: their masses, the epoch of their formation, and their evolutionary status. The properties of these first objects determine the entire subsequent evolution of the Universe, and their influence on the ionization state of diffuse gas determines the level of angular fluctuations in the cosmic background radiation. In cosmological models with cold dark matter, the formation of the first stellar objects corresponds to the compression of baryonic matter in the gravitational field of non-baryonic dark matter, which forms, in turn, so-called dark halos—condensations of dark matter in a virial state [1]. The minimum mass of the first objects can be estimated as the minimum baryonic mass that can not only sustain itself in the gravitational field of the dark halo, but also be compressed, with the subsequent formation of gravitationally bound condensations of baryons—future protostars or proto-stellar clusters.

This question has been studied by a number of groups in recent years, and the following scenario is currently generally accepted [2] (see also the reviews [3, 4]). Let us consider a mass M consisting nearly 90% of non-baryonic dark matter that separates out from the general Hubble expansion and reaches a virial state at redshift z_v. The temperature of the baryons associated with the mass M in the virial state reaches some value $T_v(M)$ that depends on M [5]:

$$T_v \cong 0.3 \left(\frac{\pi^3}{96}\right)^{1/3} \frac{G\mu m_p}{k} M^{2/3} (\Omega_m \rho_0)^{1/3} (1+z), \quad (1)$$

where $\rho_0 = 3H_0^2/8\pi G$ is the critical density. If gas at this temperature is able to cool via radiative losses in molecular hydrogen lines at the given z, it can subsequently contract and form a gravitationally bound baryonic object. Since the efficiency of the formation of H_2 molecules and cooling via their line radiation falls off with decreasing temperature, there exists at each z a minimum mass $M_m(z)$ for which the gas at the temperature T_v ends up in a cooling state over the comoving Hubble time $t_H(z)$: $M_m(z)$ is a decreasing function of z [2]. The absolute minimum mass of the first objects and the epoch of their formation in the Universe z are determined by the intersection of the mass $M_m(z)$ and the mass of the halo $M_{3\sigma}(z)$ corresponding to a fluctuation amplitude exceeding the 3σ level. According to [2], this redshift is $z = 30$, and the masses of the first objects are $M \sim 10^6 M_\odot$. Only objects with higher masses can form at lower redshifts. The cooling criterion adopted in [2] is that there be a substantial decrease in the temperature of the gas compared to the virial temperature over some specified time for each z: all objects satisfying

1063-7729/03/4712-0979$24.00 © 2003 MAIK "Nauka/Interperiodica"

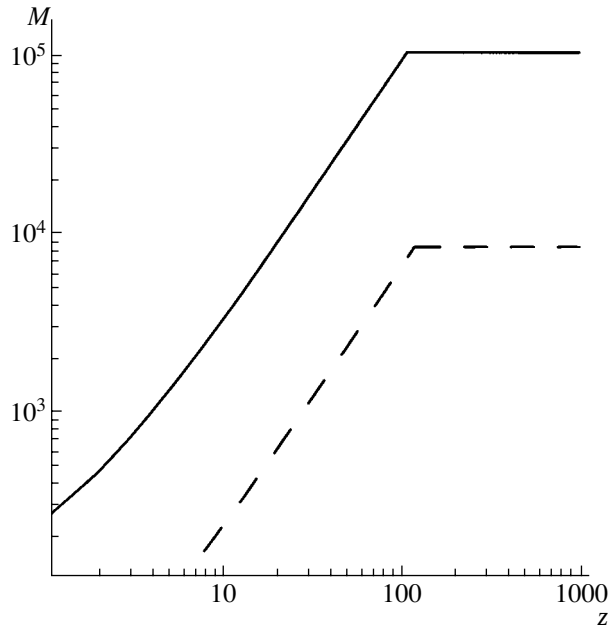

Fig. 1. Jeans mass M_J (solid curve) and the minimum mass M_c (dashed curve) in units of $M_\odot$.

the condition $T(\eta z) < \eta T(z)$ with $\eta = 0.75$ are considered to be cooling, and capable of further forming gravitationally bound baryonic systems.

Thus, this scenario is limited to rapidly evolving gas condensations that preferably form the earliest stellar systems at redshifts close to the epoch when the halo separates out from the Hubble expansion. In addition, objects with lower masses that evolve appreciably more slowly remain outside this framework, so that they become capable of forming stellar systems only at epochs close to the current epoch ($z \sim 0$), or alternatively form metal-poor gravitationally bound clumps that are captured by more massive galaxies in a bottom–up hierarchical scheme of galaxy formation.

In the current paper, we follow the evolution of such low-mass, cooling baryonic condensations over fairly long time intervals, often appreciably exceeding the comoving Hubble time. This enables us to distinguish a class of first-generation objects (i.e., those born from primordial material), whose formation extends to comparatively low redshifts, or, in some cases, even to the current epoch, so that the epochs of their formation can coincide with the formation of objects of subsequent generations, which were born from material that has already been enriched in metals. Being gravitationally bound, such objects will be essentially isolated, and will not become enriched in metals. Therefore, they could supply material for dwarf galaxies with extremely low metal abundances at the current epoch [6].

Another important circumstance of the scenario described above is that the specific process by which the dark halos are virialized is excluded from consideration, supposing only that this virialization is accomplished via violent relaxation on time scales appreciably shorter than the comoving Hubble time. (In models of violent relaxation, the characteristic virialization time for a dark halo differs from the comoving Hubble time by the factor $\sqrt{\langle\rho\rangle/\rho_h}$, where $\langle\rho\rangle$ is the mean total density in the Universe and ρ_h is the mean density of the halo.) At the same time, the process of violent relaxation proceeds with the appearance of appreciable fluctuations of the temperature and density of the relaxing system [7, 8]. In the case of a mixture of baryon and dark components, the situation is complicated by the fact that, being dissipative, the baryon component can form a structure during the virialization process; i.e., generally speaking, long before the dark-matter component reaches the virial state. A full description of this process requires three-dimensional numerical simulations of a two-fluid, self-gravitating medium, and is accordingly extremely computationally consuming. In the current paper, we propose a qualitative analysis of the chemical and thermal evolution of the baryon component based on a simple model for the virialization of dark halos. A more detailed description of the fluctuations of the potential and density, as well as the results of two-dimensional numerical computations, will be given in subsequent papers. We present here the results of computations for a model with cold, dark matter and a Λ term (a ΛCDM model) with $\Omega = 1$, $\Omega_m = 0.29$, $\Omega_b = 0.047$, $\Omega_\Lambda = 0.71$, Hubble constant $h = 0.72$, and $\sigma = 0.9$ [9], although we also carried out computations for a standard CDM model with $\Omega = \Omega_m = 1$, $\Omega_b = 0.06$, Hubble constant $h = 0.5$, and $\sigma = 0.7$. The resulting minimum masses in the CDM model exceed those in the ΛCDM model by no more than a factor of two.

Section 2 describes the model underlying our analysis of the role of shock waves in the chemical and thermal history of the baryons in the virialization process. In Section 3, we formulate the equations used to study the dynamics of colliding baryonic condensations. In Section 4, we present a simple system of equations to describe the dynamics of the halos after they reach the virial state in a "quiet" virialization model. Section 5 describes the thermal and chemical processes determining the thermal evolution of the baryons. Finally, Sections 6 and 7 present our main results and conclusions.

2. DESCRIPTION OF THE VIRIALIZATION MODEL

The simple description of the processes determining the masses of the first objects in the Universe

that we present here is based on certain assumptions about the virialization of dark-matter halos, including the character of motions within them. In this section, we present arguments supporting these assumptions, and specify the initial state of baryons in the early stages of formation of gravitationally bound objects.

2.1. Model for the Virialization of Dark Halos

The formation and virialization of massive halos can be viewed as the unification of less massive into more massive fragments (subhalos), with subsequent relaxation to a new equilibrium corresponding to the higher mass. If an even more massive halo forms via the separation of a uniform (homogeneous) mixture of baryons and dark matter, the subsequent violent relaxation is accompanied by the formation of inhomogeneities in the cold component of the dark matter due to gravitational instability, whose interactions with each other accelerate the relaxation process. Following the terminology adopted in [7], we will call this model "violent" virialization, in constrast to the "quiet" virialization considered below (see Sections 4 and 6.1). In our simple description, we will assume that, at the time of its separation from the Hubble flow and the onset of virialization, a dark-matter halo of mass M consists of an ensemble of subhalos with lower masses m_i, such that $M = \sum_i m_i$. Even if the relative velocities of the lower-mass halos were initially small, they can grow during the subsequent compression of the system as a whole, reaching values close to $v_r \sim \sqrt{GM/r}$, where r is the current radius of the halo, and thus can exceed the virial velocity $v_R = \sqrt{GM/R}$ before equilibrium is attained, where R is the radius of the system in the virial state.

2.2. Critical Mass

We restrict our consideration to those stages of evolution of the Universe when the thermal contact between matter and radiation via Compton interactions between electrons and photons had ceased to be important. This corresponds to redshifts $z < z_i$, where

$$1 + z_i = 137 \left(\frac{\Omega_b h^2}{0.022} \right)^{2/5}, \qquad (2)$$

where $h = H_0/100$ km s^{-1} Mps^{-1} is the Hubble constant and Ω_b is the baryonic density parameter. At $z > z_i$, the temperature of baryons and electrons T coincides with the temperature of the background relict radiation $T_r = 2.7(1 + z)$ K, while at $z < z_i$,

$$T_b = T_r \left(\frac{1+z}{1+z_i} \right)^2. \qquad (3)$$

In spite of the fact that the masses of dark halos are not restricted from below by gravitational instability (this restriction can be associated only with the possible cutoff of the perturbation spectrum $\delta_d(k)$), there is a critical mass below which baryons with the temperature $T_b(z)$ cannot be sustained by dark halos. We determine this critical mass by requiring that the virial temperature $T_v(M)$, which must be provided by the gravitational potential of the dark matter with mass M, be higher than the background temperature of the baryons (3): $T_v(M) > T_b(z)$. We thus obtain for the critical mass

$$M_c(z) \qquad (4)$$
$$= \begin{cases} 8.45 \times 10^3 M_\odot, & z > z_i, \\ 8.45 \times 10^3 M_\odot (1 + z/1 + z_i)^{3/2}, & z \le z_i. \end{cases}$$

Note that this value differs from the critical Jeans mass for the baryons $M_J(z)$, determined for the background baryon density and temperature. Figure 1 shows the dependences $M_c(z)$ and $M_J(z)$.

2.3. Criteria for Violent Relaxation and Virialization of the Halos

In the virial state, the outer radius of the halo is

$$R = \left(\frac{M}{24\pi^3 \Omega_m \rho_0} \right)^{1/3} (1+z)^{-1}, \qquad (5)$$

so that the characteristic velocity of the subhalos is

$$v_R = \sqrt{GM^{2/3} (24\pi^3 \Omega_m \rho_0)^{1/3} (1+z)}. \qquad (6)$$

In our model for violent relaxation, the virial state is achieved via Coulomb scatterings of the halos on each other, with the interval between scatterings being

$$t_C = \left(\frac{3\pi}{2^9} \frac{N_s r^3}{G m_i} \right)^{1/2} [\ln(N_s/\sqrt{\pi})]^{-1}. \qquad (7)$$

We will consider the system to be relaxing if t_C is less than the comoving Hubble time for a specified z and $t_C < t_{\rm H}(z)$. This corresponds to the condition

$$\frac{N_s}{\ln(N_s/\sqrt{\pi})} = \frac{2^7 \pi}{3H_0} (G\rho_0)^{1/2}. \qquad (8)$$

It is easy to see that condition (8) yields $N_s < 3$ in the ΛCDM model with $\Omega_m = 0.29$. It is reasonable to suppose that, in the simplest case, halos of cold, dark matter virialize via the development of a dipolar mode corresponding to the oncoming motion of two flows, as occurs in the theory of the evolution of adiabatic perturbations of Zeldovich [10]. Thus, as a basic violent-relaxation model, we will consider a head-on collision of two flows (two subhalos) with masses

$M_1 = M_2 = M/2$ and with velocities in the center-of-mass system equal to the virial velocities (6).

An important circumstance here is that, being collisionless, the dark matter leaves the region of shock compression of the baryons. The oncoming flows of dark matter that arise are subject to streaming instability, which can bring about a rapid relaxation of the system to the virial state, and also facilitate the fragmentation of the compressed baryonic layer.

We also investigated shocks associated with the head-on collisions of lower-mass subhalos ($m_i < M/2$). However, since these subhalos will have the same characteristic velocity (6), while the durations of the collisions between them will be shorter, $\propto m_i{}^{1/3}$, the number of H_2 molecules in the final state is smaller, and consequently the efficiency of the cooling of the gas is lower.

3. BARYONS BEHIND THE SHOCK FRONTS

When considering models for violent virialization, we restricted our analysis to a one-dimensional description of the thermal evolution of the baryons during a collision of subhalos. The legitimacy of this approach is based on the fact that, since the gas flows are subsonic behind the front, expansion in the transverse direction (parallel to the plane of the shock front) requires more time than expansion in the direction of the collision. However, we should keep in mind that the development of a strong outflow of matter in the equatorial direction (along the shock front) immediately after the collision can lower the pressure behind the front, and consequently decrease the degree of compression and the efficiency of forming molecular hydrogen. However, as a rule, this decrease in pressure does not exceed a factor of two, so that our results will provide a qualitatively correct description of the dynamics of the cooling gas in colliding subhalos. In the absence of significant thermal conductivity, the evolution of the baryons behind the shock front can be described by a system of ordinary differential equations written for an individual Lagrange element of the fluid. We will assume that the temperatures of the electrons and heavy particles (atomic hydrogen and helium and their ions) coincide behind the front. In this case, the system of dynamical equations describing the thermal properties and chemical transformations of the baryon component behind the shock front during a collision of subhalos takes the form

$$\dot{T} = -\sum_i \Lambda_i(T, x_2, x, n), \tag{9}$$

$$\dot{x} = I(T, x, n) - R(T, x, n), \tag{10}$$

$$\dot{x}_2 = F(T, x, n) - D(T, x, n), \tag{11}$$

where Λ_i is the cooling rate of the gas associated with process i, I and R are the ionization and recombination rates for hydrogen (we neglect here the contribution of helium to thermal and ionizational processes), F and D are the total rates of the formation and dissociation of molecular hydrogen, x is the degree of ionization, x_2 is the relative number density of H_2, and T is the temperature of the gas. We assume that each gas element behind the front evolves isobarically, and describe the density using the expression

$$n = \frac{p}{\mu k T}, \tag{12}$$

$\mu = \rho/n m_p = 1.4$. Here, we have neglected the contribution of electrons, protons, and other particles taking part in the chemical and ionization equilibrium behind the front, since their relative number densities always remain substantially lower than unity. The pressure p is determined from the condition of conservation of momentum at the discontinuity:

$$p = \frac{15}{9}\rho_i v^2 \tag{13}$$

for an adiabatic shock, where v is the velocity of the cloud relative to the center of mass, ρ_i is the density of baryons in the subhalos before their collision, and

$$p = \rho_i v^2 \tag{14}$$

for a radiative shock [11]. We adopt for the initial temperature the temperature of the gas behind the shock front:

$$T_0 = \frac{15}{36}\frac{m_p}{k} v^2 \tag{15}$$

for an adiabatic shock and

$$T_0 = \frac{m_p}{4k} v^2 \tag{16}$$

for a radiative shock. The initial degree of ionization and number density of H_2 were taken to be equal to their cosmological values.

4. BARYONS IN THE QUIET VIRIALIZATION MODEL

The system of dynamical equations that we will use to describe the dynamics of the baryons in the field of the dark-matter halo in the case of quiet (spherically symmetrical) virialization can be written

$$\dot{R} = u, \tag{17}$$

$$\dot{u} = \frac{5}{11}\left[\frac{4kT}{\mu m_p R} - \frac{4}{3}\pi G(\rho_d + \rho)R\right], \tag{18}$$

$$\dot{T} = -\frac{2m_p}{3k}\left[\frac{3kT}{\mu m_p R} + \frac{4}{5}\pi G(\rho_d + \rho)R\right]u - \frac{2L}{3k\mu n}, \tag{19}$$

where $n = \rho/\mu m_p$, G is the gravitational constant, R the radius of the cloud, and u the velocity of the cloud particles at its boundary. The system (17)–(19) is obtained from the integral form of the gas-dynamical equations [12]

$$\int_V \rho \mathbf{v} dV + \int_{t_0}^{t} dt \int_{\mathbf{S}} \rho v^2 d\mathbf{S} \qquad (20)$$
$$= -\int_{t_0}^{t} dt \int_{\mathbf{S}} p d\mathbf{S} + \int_{t_0}^{t} dt \int_V \mathbf{F} dV,$$

$$\int_V \rho(\epsilon + v^2/2) dV - \int_{V_0} \rho(\epsilon + v^2/2) dV \qquad (21)$$
$$= -\int_{t_0}^{t} dt \int_{\mathbf{S}} p\mathbf{v} d\mathbf{S} - \int_{t_0}^{t} dt \int_V L dV + \int_{t_0}^{t} dt \int_V \mathbf{F}\mathbf{v} dV.$$

For a spherical cloud, substituting the thermodynamical quantities in the integrals with their mean values and assuming that the local velocity of the compression is proportional to the local radius (which, strictly speaking, is valid for a free-fall regime), $v(r) = -\alpha r$. In these equations, L [erg cm^{-3} s^{-1}] is the radiative-cooling rate per unit volume, $F(r) = -GM\rho/r^2$, $p = knT$, and $\epsilon = 3kT/2\mu m_p$. In the system made up of (17)–(19) and (20)–(21), it is assumed that the kinetic energy of the compressed gas is transformed into thermal energy on short timescales. It can readily be seen from (18)–(19) that, in the absence of radiative cooling, $L \equiv 0$, the gas temperature increases to the virial temperature as it is compressed, T_v (1), after which the rate goes to zero, and the compression ceases.

We solved the system (17)–(19) for a model with quiet evolution of the halo, without including the effects of fluctuations and the appearance of shocks. We considered the evolution of the baryon component from the time the dark halo separated out; i.e., from the turning point when the Hubble expansion within some perturbed region of dark matter is replaced by compression. We described the evolution of the dark matter using the approximate expression given in [2], which reproduces well the exact solution. We used the temperature and density corresponding to the turning point as their initial values, also taking the initial densities of electrons and H_2 molecules to be equal to their cosmological values at that time. The system of dynamical equations for the baryons (17)–(19) was solved further for the case of a specified (approximated) dark-matter background.

Rates of chemical reactions

Reaction	Reaction coefficient k, cm^3 s^{-1}	Ref.
$H^+ + e^- \to H + h\nu$	$k_1 \approx 1.88 \times 10^{-10} T^{-0.64}$	[16]
$H + e^- \to H^- + h\nu$	$k_2 \approx 1.83 \times 10^{-18} T^{0.88}$	[16]
$H^- + H \to H_2 + e^-$	$k_3 \approx 1.3 \times 10^{-9}$	[17]
$H^+ + H \to H_2^+ + h\nu$	$k_5 \approx 1.85 \times 10^{-23} T^{1.8}$	[18]
$H_2^+ + H \to H_2 + H^+$	$k_6 \approx 6.4 \times 10^{-10}$	[19]
$D^+ + H_2 \to HD + H^+$	$\alpha_1 \approx 2.1 \times 10^{-9}$	[15]
$H^+ + HD \to H_2 + D^+$	$\alpha_2 \approx \alpha_1 e^{-465K/T}/4$	[15]

5. THERMAL AND CHEMICAL PROCESSES

The main cooling agents for the baryons in the redshift range of interest to us ($z < 200$) are molecular hydrogen, H_2, and HD. In the computations, we adopted for the cooling function of the H_2 molecules [13] the expression

$$L_{H_2} \simeq \frac{{L_r}^E}{1 + n_{cr}/n}, \qquad (22)$$

where the critical density is

$$n_{cr} \equiv \frac{{L_r}^E}{{L_r}^{(n\to 0)}} n, \qquad (23)$$

which depends only on the temperature, not on n. The cooling rate in LTE is

$${L_r}^E \simeq \frac{1}{n}\left\{\frac{9.5 \times 10^{-22} {T_3}^{3.76} e^{-(0.13/T_3)^3}}{1 + 0.12 {T_3}^{2.1}} + 3 \times 10^{-24} e^{-0.51/T_3}\right\} \text{ erg cm}^3/\text{s}, \qquad (24)$$

while the cooling rate in the limit of low densities is

$${L_r}^{(n\to 0)} \simeq \frac{5}{4}\gamma_2 (E_2 - E_0) e^{-(E_2 - E_0)/kT} + \frac{7}{4}\gamma_3 (E_3 - E_1) e^{-(E_3 - E_1)/kT}. \qquad (25)$$

Here, $T_3 = T/1000\ K$ and $E_J = J(J+1)E_1/2$, where $E_1/k \simeq 171\ K$, and γ_2 and γ_3 are the collisional de-excitation rates for transitions from the $J = 2$ and $J = 3$ rotational levels [13]:

$$\gamma_J = \left(\frac{10^{-11} {T_3}^{1/2}}{1 + 60 {T_3}^{-4}} + 10^{-12} T_3\right) \times \left\{0.33 + 0.9 \exp\left[-\left(\frac{J - 3.5}{0.9}\right)^2\right]\right\} \text{ cm}^3/\text{s}. \qquad (26)$$

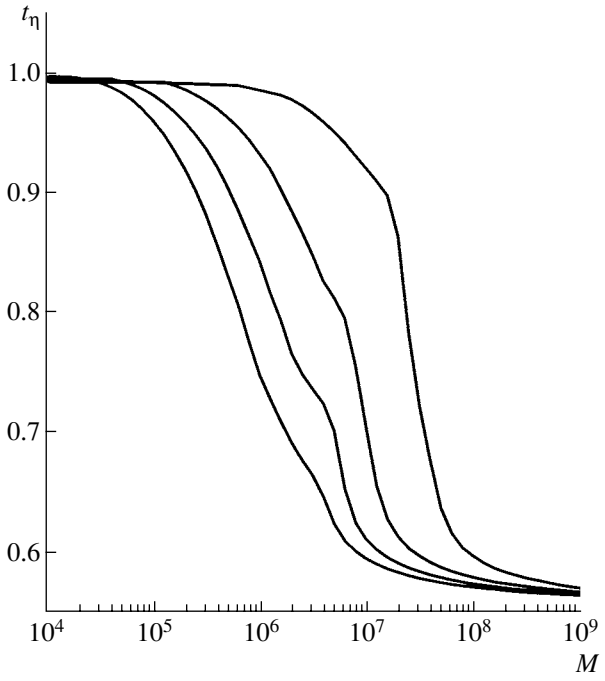

Fig. 2. The $t_\eta(M)$ relation, where t_η is the time that has passed from the separation of the halo from the cosmological expansion to the collapse ($\eta = 1$) in units of the comoving Hubble time $t_{\rm H}(z)$, for $z_v = 40...10$ in steps of ten (left to right). For low masses, this time is comparable to the current age of the Universe. The masses are indicated in units of $M_\odot$.

Equation (25) assumes that the ratio of the ortho and para configurations of H_2 is 3 : 1. The first and second terms give the contributions to the cooling from para- and ortho-H_2. It is shown in [14] that, in the volume reactions

$$H_2(\text{orto}) + H^+ \rightarrow H_2(\text{para}) + H^+, \qquad (27)$$

all the ortho-H_2 is transformed into para-H_2. Thus, the corresponding cooling rate is obtained by multiplying the first term of (25) by four; i.e., Eq. (25) is replaced by the simpler expression

$$L_r{}^{(n\rightarrow 0)} \simeq 5\gamma_2(E_2 - E_0)e^{-(E_2-E_0)/kT}. \qquad (28)$$

For the rate of cooling by the HD molecules, we adopted the expression [13]

$$L_{\rm HD} = 6.6 \times 10^{-23} T^{-1/2} \qquad (29)$$
$$\times \exp\left(-\frac{128.4\text{K}}{T}\right) \text{ erg cm}^3/\text{s}.$$

The critical density for the HD molecules is substantially higher than the values attained in our computations. The computations included the chain of reactions presented in the table, which are the dominant processes under the conditions considered. A very complete analysis of the reactions in a primary gas is presented in [15]. We neglected the contribution of the photodissociation of H^- and H_2^+ ions by photons of the background radiation, since their energies are two to three orders of magnitude lower than the binding energies of these ions in the redshift range considered, $z \leq 40$.

Taking into account the fact that the reactions $D^+ + H_2 \longrightarrow HD + H^+$ and $H^+ + HD \longrightarrow H_2 + D^+$ are rapid, i.e., their characteristic time scales are substantially shorter than other time scales in the problem, the system of equations describing the relative densities of electrons $x = n_e/n$ and H_2 molecules, $f = 2n(H_2)/n$, can be written

$$\dot{x} = -k_1 n x^2, \qquad (30)$$

$$\dot{f} = 2k_m n(1 - x - f)x, \qquad (31)$$

where, in accordance with [2] and allowing for the fact that photodissociation of the H^- and H_2^+ ions by cosmic background photons is negligible,

$$k_m = k_2 + k_5 \qquad (32)$$

(the first and second terms describe the formation of H_2 in reaction chains involving the H^- ion and H^+ ion, respectively). The collisional ionization of H atoms and dissociation of H_2 are negligibly small under the low-temperature conditions we are considering.

At temperatures below $\simeq 400$ K, the effect of the chemical fractionation of deuterated molecular hydrogen becomes important [20]. In late stages of the compression of the condensations as the gas is cooled to a temperature of ~ 30 K, all the deuterium may be transformed into molecular form [21]. However, under the conditions for the initial evolution of the halos considered here, when the temperature of the baryons is not able to fall below 30–100 K, the contribution of HD molecules to the thermal balance is only 5–10%.

6. RESULTS

6.1. Model for Quiet Virialization

At each fixed redshift, we carried out computations for grids of halo masses exceeding the critical mass $M_c(z)$, and determined the minimum mass $M_m(z)$ for which the baryons are able to be cooled and compressed. The computations were conducted for a standard CDM model and for a CDM model with a Λ term. The standard CDM model gives minimum masses that exceed their values for the ΛCDM model by no more than a factor of two. We fixed the baryon mass, taking it to be equal to $(\Omega_b/\Omega_m)M$, where M is the total mass of the cloud, including the baryon and dark matter components. Following the standard approach (see, for example, [2, 22]), we suppose that,

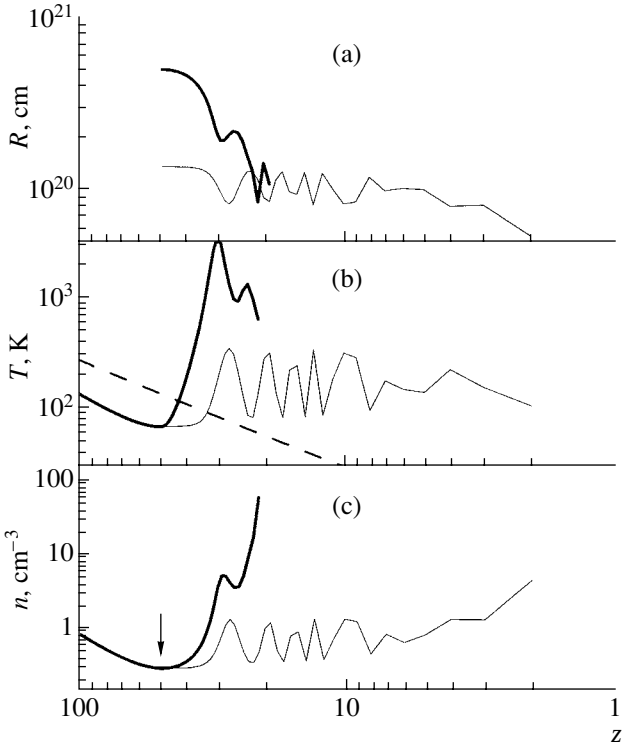

Fig. 3. Evolution of the (a) radius, (b) temperature, and (c) density for two types of halos (with masses close to and appreciably exceeding the critical mass, shown by the dotted and solid curves, respectively) that have become virialized at $z_v = 30$. The time of the separation from the cosmological expansion is indicated by the arrow. The middle graph also shows the temperature of the cosmic background radiation (dashed curve).

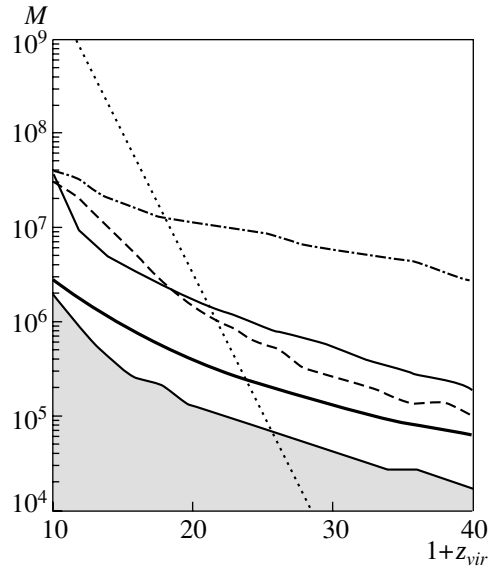

Fig. 4. Minimum mass required for collapse. Only masses lying above the shaded region can be compressed and form objects by the current age of the Universe (slowly evolving objects). Masses lying above the thin, solid curve can be compressed and form objects over the local (at the time of virialization) Hubble time (rapidly evolving objects corresponding to the compression criterion [2]). The thin, dot–dash curve corresponds to rapidly evolving objects satisfying our compression criterion. The thin, dotted curve corresponds to objects forming by the epoch $z = 6$. The bold, solid curve shows the minimum masses for rapidly evolving objects in a violent-virialization scenario. The dashed curve corresponds to 3σ fluctuations in the ΛCDM model ($\sigma_8 = 0.9$). The masses are indicated in units of $M_\odot$.

at some z, a perturbed region separates out from the general Hubble expansion, begins to be compressed, and, at $z \simeq z_v$, the halo of dark matter relaxes to the virial state with characteristic density

$$\rho_v \approx 18\pi^2 \rho_0 \Omega_d (1+z)^3, \qquad (33)$$

where $\Omega_d = \Omega_m - \Omega_b$. We used the following approximation to describe the transition to the virial state in the computations [2]:

$$\rho(z) \approx \rho_0 \Omega_d (1+z)^3 \exp\left(\frac{1.9A}{1-0.75A^2}\right), \qquad (34)$$

where

$$A(z) \equiv \frac{1+z_v}{1+z}. \qquad (35)$$

We assumed that the density of baryons in the halo varies proportional to the density of dark matter before the separation from the Hubble expansion, after which we found the solution to the system of equations (17)–(19). Thus, in contrast to [2], we separate the evolution of the baryon component from the non-baryon component at stages following the separation from the cosmological expansion. As a criterion for the formation of a gravitationally bound baryonic object, we adopted the condition that, in the process of the cooling and compression of the baryon component in the total gravitational field of the baryons and dark matter, the density of baryons within the radius occupied by the baryon component must exceed the density of the non-baryonic dark matter by a specified factor $\eta > 1$. This is a stricter condition than the condition that the temperature of the baryons decrease by a factor of 1.5, as was adopted in [2]. Figure 2 shows the $t_\eta(M)$ dependence, where t_η is the time that has passed from the separation from the cosmological expansion to the final compression (when $\eta > 1$) in units of the comoving Hubble time $t_{\rm H}(z)$ for various z_v.

Figure 3 depicts the evolution of the radius, temperature, and density for two values of z_v for two types of halos: those with masses close to and much greater than the critical mass $M_c(z)$. There are obvious differences in the evolution of these halos. In the former case, molecular hydrogen forms slowly in the com-

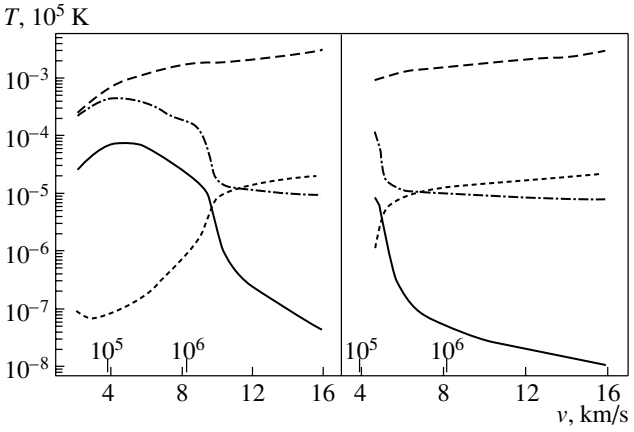

Fig. 5. Dependences of the number densities of electrons, molecular hydrogen, and HD and of the baryon temperature (solid, dashed, dotted, and dot–dash curves, respectively) on the halo collision velocity for head-on collisions. The masses of the halos are $M_s = 10^4 M_\odot$ (left plot) and $M_s = 10^5 M_\odot$ (right plot). The marks below indicate the virial velocities corresponding to the indicated series of masses (in units of $M_\odot$).

pressed baryon component, so that only after several oscillations does enough H_2 accumulate to provide the subsequent cooling and monotonic compression of the baryons. In the latter case, a sufficient quantity of H_2 forms even in the initial stages of the compression, giving rise to the rapid cooling and compression of the baryon component. For each z_v, we computed the dynamics of the compression of an ensemble of halos with various masses, in order to find the critical (minimum) mass $M_m(z)$ of a halo that is able to be compressed over the local (i.e., modern) Hubble time. These masses lie above the shaded region in Fig. 4, which also shows the minimum mass for rapidly (over the comoving Hubble time) evolving halos for the criterion applied in [2] (solid, thin curve; the computations in [2] were carried out for the standard CDM model, while we have recomputed the results of [2] for a ΛCDM model), and for our stricter criterion (thin, dot–dash curve; corresponding to the rapid compression of the baryons to $M_b/M_d > 1$ over the comoving Hubble time). We can see that the differences in the masses for rapidly and slowly evolving halos can reach two orders of magnitude. The thin dashed curve corresponds to masses that are compressed and form gravitationally bound baryonic condensations by the epoch $z = 6$ in our model. The upper limit for the mass of the baryonic objects being born is determined by the curve $M_{3\sigma}$, which corresponds to exceeding the density fluctuations at the 3σ level [4] (dashed curve in Fig. 4). Thus, gravitationally bound objects with a wide range of masses $M_m(z) < M < M_{3\sigma}(z)$ are simultaneously born and evolve in the Universe. The more massive objects evolve more rapidly, while lower-mass objects can give rise to new stellar systems only at later epochs.

6.2. Model for Violent Virialization

In accordance with the above, we also carried out computations for each fixed redshift for a grid of halo masses M in this model. The halos relaxed via the collision of two subhalos with half the total mass, $M_s = M/2$, and with the virial velocities in the center-of-mass system:

$$v_R \cong 1.16 \times 10^5 \left(\frac{M}{10^6\ M_\odot}\right)^{1/3} (1+z)^{1/2}\ \text{cm/s}. \quad (36)$$

The computations were conducted for collisional velocities from 2 to 16 km/s, which corresponds to the virial velocities for halo masses $M = (2 \times 10^4 – 8 \times 10^6) M_\odot$ at redshift $z = 30$.

A halo was considered to be capable of subsequently forming a stellar system if the Jeans mass in the cooled gas after the collision was smaller than the total mass of the baryons. The corresponding halo mass is shown by the bold solid curve in Fig. 4. The computations were carried out over times equal to one dynamical time for the colliding subhalos: $t_d = R/v$. This corresponds to the fact that the compressed object will begin to expand and the formation of H_2 molecules will cease at $t > t_d$. Figure 5 presents the dependence of the number densities of the electrons, H_2, and HD and of the temperature on the velocity of the subhalo collisions at redshift $z = 30$, for $M_s = 10^4 M_\odot$ (left) and $M_s = 10^5 M_\odot$ (right). The mass of a halo (in $M_\odot$) for which the virial velocity is equal to the corresponding velocity on the x axis is shown below.

We expect that, in a more complex scenario when the relaxation of the halos occurs due to interactions between multiple (more than two), less massive subhalos, the formation of H_2 molecules and the cooling due to them will be less efficient than in the case considered above, since the times for their interactions will be shorter, $t_d \propto m_i^{1/3}$. Note, however, that the final count at time $t = t_d$ corresponds to a lower limit to the effects discussed, since, after a collision, the expanding gas fragments take part in subsequent collisions with fragments from other collisions, with the characteristic relative velocities being $\sqrt{2} v_R$. As a result, a random, supersonic velocity and density field that is close to a logarithmic normal distribution will be established in the system, as follows from numerical simulations [23]. One consequence of this will be the ongoing, more efficient formation of H_2 molecules behind the shock front, enhancing the overall cooling of the baryon component of the halo.

6.3. Survival of Low-Mass Objects

A fundamental question is the survival of slowly evolving objects in the field of the ionizing radiation created by the stellar population born from rapidly evolving systems with high masses (above the limit of [2], shown by the thin solid curve in Fig. 4). One of the most important consequences of the formation of the first objects is the secondary ionization and heating of the Universe [4, 9, 24–27]. The subsequent evolution of protogalaxies and the efficiency of star formation in them can be substantially affected by the ionizing radiation emitted by the first objects formed. We note in this connection that the most recent observations of distant quasars with redshifts $z \gtrsim 4.5$ as part of the Sloan Digital Sky Survey appear to indicate that the secondary ionization and heating occurred near redshifts $z \sim 6$, so that the flux of ionizing radiation drops sharply at high redshifts ($z > 6$), as is testified to by the latest observations of the Gunn–Peterson effect [28–30] and the WMAP data [9]. This may indicate that the very first objects in the Universe formed low-mass stars that were not able to make an appreciable contribution to ionizing radiation, that the bulk of the ionizing photons are trapped in dense zones of HII, or alternatively that the number of these earliest objects was small. In any case, this means that low-mass (slowly evolving) objects with masses in the interval indicated above can evolve without being appreciably affected by ionizing radiation, at least to redshifts $z \sim 6$. Indeed, at the densities that are reached in these objects at stages when the central baryon density exceeds the density of the non-baryonic dark component by an order of magnitude, the number density of hydrogen is $n \sim 10\ \mathrm{cm}^{-3}$. At ionization rates $\xi_{\mathrm{H}} < 4 \times 10^{-14}\ \mathrm{s}^{-1}$, corresponding to redshifts $z > 6$ [30], the degree of ionization brought about by external radiation is only $x < 4 \times 10^{-2}$, and the ionization induced heating rate is $\Gamma \lesssim 6 \times 10^{-26}\ \mathrm{erg\ s}^{-1}$, while the rate of cooling by molecular hydrogen at these stages lies in the range $\sim (4\text{–}6) \times 10^{-26}$ for masses of $(0.3\text{–}3) \times 10^6 M_\odot$, which virialize at $z = 30\text{–}20$. The rate of dissociation of H^- ions by ionizing radiation of the Lyman continuum for the flux estimated in [30] is $\sim 4 \times 10^{-18}\ \mathrm{s}^{-1}$, substantially lower than the rate at which they are transformed into H_2 ($\sim 10^{-8}\ \mathrm{s}^{-1}$). Note that, under these conditions, the ionizing radiation can stimulate the formation of H_2 if it leads to an enhancement of the electron density to $x \sim 10^{-2}$. In the more realistic case of violent relaxation of the halos with the formation of multiple shocks, the efficiency of the secondary ionization and photodissociation of H_2 decreases, because the density of gas in regions through which the shocks have passed is higher than the average density. Thus, we expect that some low-mass halos can be preserved in the external field of ionizing radiation, and, moreover, some fraction of these can subsequently also contribute to this field. It will be possible to draw firm conclusions only on the basis of self-consistent computations of the formation of the first stellar objects and the production of ionizing radiation by them.

7. CONCLUSIONS

(1) We have considered the evolution of low-mass, slowly evolving first objects—so-called dark-matter halos—which give rise to the first stellar population of the Universe. We showed in our simple (quiet) virialization model that baryonic objects with masses exceeding $(0.3\text{–}10) \times 10^5 \Omega_b M_\odot$ can be compressed and give rise to stellar systems at the current epoch. Baryonic objects with masses $(0.1\text{–}12) \times 10^6 \Omega_b M_\odot$, which virialize at $z = 40\text{–}10$, form gravitationally bound systems by $z = 6$. Since the mass function of the first objects decreases with the mass (probably more rapidly than M^{-1}), this may mean that the production of ionizing photons at $z < 6$ will be higher than is usually predicted.

(2) We have shown that the estimated minimum mass of the forming objects depends substantially on the adopted baryon-cooling criterion. In particular, if we take such objects to be those in which the baryons are compressed during the cooling process to densities exceeding the density of dark matter in the central regions of the halo, then this state can be reached over the comoving Hubble time only by objects whose masses exceed the known limit [2] by half an order of magnitude.

(3) We have examined the influence of the shock waves that necessarily arise during the virialization of the halos on the thermal regime of the baryons in a simplified approach. The minimum baryon temperature is achieved during collisions of fragments with masses that are half the mass of the halo. Although the fraction of such halos is probably small, precisely they give rise to the formation of the least massive first objects, with masses $(0.6\text{–}12) \times 10^5 M_\odot$, which is lower than the limit of [2] by a factor of 2–12.

ACKNOWLEDGMENTS

The authors thank the referee for comments. This work was supported by the Russian Foundation for Basic Research (project no. 00-02-17689) and the State Science and Technology Program "Astronomy" (project 1.3.1.1).

REFERENCES

1. M. J. Rees and J. P. Ostriker, Mon. Not. R. Astron. Soc. **179**, 541 (1977).
2. M. Tegmark, J. Silk, M. J. Rees, *et al.*, Astrophys. J. **474**, 1 (1997).
3. D. Galli and F. Palla, *H_2 in Space*, Ed. by F. Combes and G. Pineau des Forêts (Cambridge University Press, Cambridge, 1999), p. 41.
4. R. Barkana and A. Loeb, Phys. Rep. **349**, 125 (2001).
5. A. Blanchard, D. Valls-Gabaud, and G. A. Mamon, Astron. Astrophys. **264**, 365 (1992).
6. T. X. Thuan, Yu. I. Izotov, and V. A. Lipovetsky, Astrophys. J. **477**, 661 (1997).
7. D. Lynden-Bell, Mon. Not. R. Astron. Soc. **136**, 101 (1967).
8. F. H. Shu, Astrophys. J. **225**, 83 (1978).
9. D. N. Spergel, L. Verde, H. V. Peiris, *et al.*, Astrophys. J. (2003, in press); astro-ph/0302209.
10. Ya. B. Zel'dovich, Astrofizika **6**, 119 (1970).
11. A. A. Suchkov, Yu. A. Shchekinov, and M. A. Edel'man, Astrofizika **18**, 629 (1982).
12. L. D. Landau and E. M. Lifshitz, *Hydromechanics* (Nauka, Moscow, 1986).
13. D. Hollenbach and C. F. McKee, Astrophys. J., Suppl. Ser. **41**, 555 (1979).
14. T. Abel, P. Anninos, Y. Zhang, and M. L. Norman, New Astron. **2**, 181 (1997).
15. D. Galli and F. Palla, Astron. Astrophys. **335**, 403 (1998).
16. J. B. Hutchins, Astrophys. J. **205**, 103 (1976).
17. T. Hirasawa, Progr. Theor. Phys. **42**, 523 (1969).
18. P. R. Shapiro and H. Kang, Astrophys. J. **318**, 32 (1987).
19. Z. Karpas, V. Anicich, and W. T. Huntress, J. Chem. Phys. **70**, 2877 (1979).
20. V. K. Khersonskiĭ and D. A. Varshalovich, Pis'ma Astron. Zh. **3**, 506 (1977) [Sov. Astron. Lett. **3**, 277 (1977)].
21. Yu. A. Shchekinov, Pis'ma Astron. Zh. **12**, 499 (1986) [Sov. Astron. Lett. **12**, 211 (1986)].
22. P. J. E. Peebles, *Principles of Physical Cosmology* (Princeton Univ. Press, Princeton, 1993).
23. P. Padoan, R. Jimenez, and B. Jones, Mon. Not. R. Astron. Soc. **285**, 711 (1997).
24. T. Abel and Z. Haiman, *H_2 in Space*, Ed. by F. Combes and G. Pineau des Forêts (Cambridge University Press, Cambridge, 1999), p. 237.
25. Z. Haiman and A. Loeb, Astrophys. J. **483**, 21 (1997).
26. A. Ferrara, Astrophys. J. **499**, L17 (1998).
27. B. Ciardi, A. Ferrara, and T. Abel, Astrophys. J. **533**, 594 (2000).
28. S. G. Djorgovski, S. Castro, D. Stern, and A. A. Mahabal, Astrophys. J. **560**, L5 (2001).
29. X. Fan, V. K. Narayanan, M. A. Strauss, *et al.*, Astron. J. **123**, 1247 (2002).
30. A. Lidz, L. Hui, M. Zaldarriaga, and R. Scoccimarro, Astrophys. J. **579**, 491 (2002).

Translated by D. Gabuzda

Astronomy Reports, Vol. 47, No. 12, 2003, pp. 989–999. Translated from Astronomicheskiĭ Zhurnal, Vol. 80, No. 12, 2003, pp. 1075–1085.
Original Russian Text Copyright © 2003 by Postnov, Cherepashchuk.

Masses of Stellar Black Holes and Testing Theories of Gravitation

K. A. Postnov and A. M. Cherepashchuk

Sternberg Astronomical Institute, Moscow State University, Universitetskiĭ pr. 13, Moscow, 119899 Russia

Received April 21, 2003; in final form, July 26, 2003

Abstract—The paper analyzes the mass distribution of stellar black holes derived from the light and radial-velocity curves of optical stars in close binary systems using dynamical methods. The systematic errors inherent in this approach are discussed. These are associated primarily with uncertainties in models for the contribution from gaseous structures to the optical brightness of the systems under consideration. The mass distribution is nearly flat in the range 4–15 $M_\odot$. This is compared with the mass distribution for black holes in massive close binaries, which can be manifest as ultrabright X-ray sources ($L_x > 10^{39}$ erg/s) observed in other galaxies. If the X-ray luminosities of these objects correspond to the Eddington limit, the black-hole mass distribution should be described by a power law, which is incompatible with the flat shape derived dynamically from observations of close binaries in our Galaxy. One possible explanation of this discrepancy is the rapid evaporation of stellar-mass black holes predicted in recent multi-dimensional models of gravity. This hypothesis can be verified by refining the stellar black-hole mass spectrum or finding isolated or binary black holes with masses below $\sim 3M_\odot$. © *2003 MAIK "Nauka/Interperiodica".*

1. INTRODUCTION

The discovery of approximately twenty black holes with stellar masses and about one hundred supermassive black holes (e.g. reviews [1, 2]) raises the question of their demography; i.e., the relationship of these extreme objects to other objects in the Universe, as well as to the deep physical properties of space–time. Bimodality in the mass distribution for stellar-mass relativistic objects was recently detected [3–5]. The masses of neutron stars lie in the narrow range $M_{NS} = (1–2)M_\odot$, with the mean value being $(1.35 \pm 0.15)M_\odot$. On the other hand, the masses of black holes are distributed over the fairly broad range $M = (4–15)M_\odot$, with the mean value being $8–10M_\odot$. No neutron stars or black holes have been found in the mass interval $2–4M_\odot$, despite the fact that masses have now been measured for almost 40 relativistic objects. This gap in the mass distribution of relativistic objects at masses of $2–4M_\odot$ cannot be due to observational selection effects [2–5], and seems especially surprising from the viewpoint of recent data on the mass distribution of the CO cores of Wolf–Rayet stars at the end of their evolution [5], which cover the wide range $M^f_{CO} = (1-2)–(20-44)M_\odot$ and are distributed continuously. Since Wolf–Rayet stars in close binary systems are commonly thought to be the progenitors of relativistic objects [6–8], the large difference between the distribution of the final masses of the CO cores of Wolf–Rayet stars and the masses of the resulting relativistic objects is a non-trivial observational fact, which must be explained. The bimodal mass distribution for relativistic objects was interpreted in [9, 10] in terms of modern concepts about the late stages of stellar evolution and explosions of collapsing supernovae (Types II and Ib, c). As an alternative way to explain the broad distribution of massses for stellar-mass black holes and the lower limit observed in binary systems, $\sim 4M_\odot$, we consider here modern, multi-dimensional theories of gravity, which enable us to view the mechanism and characteristic time for the quantum evaporation of black holes in a new light.

2. METHODS FOR DETERMINING BLACK-HOLE MASSES IN BINARY SYSTEMS

It is important to answer the question of whether the observed broad distribution of black-hole masses in the interval $4–15M_\odot$ is real, or whether these masses are actually distributed in accordance with some other law (for example, concentrated in a narrower interval), with the observed scatter being due to errors in the derived masses.

Most of the information about the mass of a black hole in a binary system in which the secondary is an optical star is contained in the mass function of the optical star [2, 11]:

$$f_v(m) = \frac{m_x^3 \sin^3 i}{(m_x + m_v)^2} = 1.038 \times 10^{-7} K_v^3 P (1-e^2)^{3/2}, \quad (1)$$

1063-7729/03/4712-0989$24.00 © 2003 MAIK "Nauka/Interperiodica"

which is derived from its radial-velocity curve; the optical star is considered to be a point mass moving along a Keplerian ellipse. Here, m_x and m_v are the masses of the black hole and optical star in solar units, K_v the semi-amplitude of the radial-velocity curve of the optical star (in km/s), P the orbital period of the binary system (in days), and e the eccentricity of its orbit. In reality, the optical star is not a point mass, since its shape is disturbed by tidal interactions with the black hole and its atmosphere is heated by X-ray radiation from the black-hole accretion disk. Taking these effects into account shows that they affect the derived black-hole mass most when the component-mass ratio is $q = m_x/m_v < 1$ [11]. In this case, the center of mass of the binary system is located inside the optical star (as occurs, for example, in the systems Cyg X-1, LMC X-1, and SS 433, in which $q = 0.3$–0.6), and the distortion of the spectral-line profiles used to derive the radial-velocity curve is greatest. When $q = 0.3$–0.6, corrections of the mass function $f_v(m)$ for the effects of the finite size of the optical star do not exceed 10%, and can be reliably estimated using modern methods for synthesizing the line profiles and radial-velocity curves of X-ray binaries [11–13].

In the case of X-ray binaries with massive (O–B) companions (Cyg X-1, LMC X-1, LMC X-3, and SS 433), there is another effect that disturbs the line profiles and radial-velocity curve of the optical star: the variable (depending on the phase of the orbital period) selective absorption of the light of the optical star by its intense stellar wind. (The mass-loss rates of such stars are typically $\sim 10^{-6}$–$10^{-7} M_\odot$/yr, and reach $10^{-4} M_\odot$/yr in the case of SS 433.) The absorption coefficient at the line center is considerably greater than in the neighboring continuum. Therefore, the central part of the absorption line is formed in the upper layers of the stellar atmosphere, at the base of the stellar wind, where the radial velocity of the plasma outflow reaches a few tens of km/s. Since the free-fall acceleration in a star with an almost filled Roche lobe varies over the stellar surface, the velocity and intensity of the wind near its base will also vary over the stellar surface, resulting in additional orbital-phase-dependent Doppler shifts of the absorption lines in the spectrum of the optical star and distortion of its radial-velocity curve [14]. Moreover, in the case of an X-ray binary with an elliptical orbit, nonradial pulsations can be excited in the optical star, as occurs in the system containing the neutron star Vela X-1 [15]. This also results in additional distortion of the radial-velocity curve of the optical star and leads to systematic errors in the derived mass of the relativistic object.

In the case of large mass ratios $q > 1$, the center of mass of the system is located outside the body of the optical star, and the effect of the finite size of the optical star becomes small. This is especially important because the masses of 15 of 18 black holes were determined in transient, low-mass X-ray binaries (X-ray novae with large mass ratios, $q > 1.5$). Therefore, the masses of most of the black holes are affected only slightly by the finite sizes of their optical components. Since the stellar winds from the low-mass (A–M) stars that are the companions of the black holes in X-ray novae are weak, the effect of selective absorption of the light of the optical stars by their wind is also small. The orbits of all X-ray novae with low-mass (A–M) companions are circular, and the optical stars in these systems fill their Roche lobes.

The mass of the invisible companion (black hole) in a binary system is derived from the mass function of the optical star $f_v(m)$:

$$m_x = f_v(m)\left(1+\frac{1}{q}\right)^2 \frac{1}{\sin^3 i}. \qquad (2)$$

The uncertainty in the black-hole mass includes random and systematic errors. The random errors can be reduced by increasing the accuracy and duration of the observations. The systematic errors are due to uncertainty in the model for the X-ray binary. Taking the systematic errors into account when determining black-hole masses is very difficult. Let us consider the influence of systematic errors in the parameters q, i on the corresponding estimate of the black-hole mass m_x.

The parameter q is usually estimated from the rotational broadening of absorption lines in the spectrum of the optical star. In most close X-ray binaries containing black holes (in particular, in X-ray novae), the optical star fills its Roche lobe, whose relative size depends on the mass ratio q. On the other hand, the larger the absolute size of the optical star, the greater the rotational broadening of absorption lines in its spectrum. As a result, assuming the axial and orbital rotations are synchronized, we obtain the following equation determining q [1, 11–13]:

$$v \sin i = 0.462 K_v q^{-1/3}\left(1+\frac{1}{q}\right)^{2/3}. \qquad (3)$$

The rotational broadening $v \sin i$ varies with the phase of the orbital period, since the dimensions of the star along the line connecting the component centers are different from those perpendicular to this direction [11]. In addition, X-ray heating of the optical star gives rise to an emission component in the lines that depends on the phase of the orbital period, which distorts the standard absorption-line profile [13]. The

corresponding errors in $v \sin i$ values derived from the analysis of absorption-line profiles in the spectrum of the optical star can be as large as 10–20%. A new method for determining q and i from the orbital variability of the absorption-line profiles in the optical spectra of close X-ray binaries was suggested in [16, 17]. Modern methods for synthesizing line profiles in the spectra of the optical components of close X-ray binaries can take into account the spectral variability of the optical star and thereby reduce systematic errors as much as possible. We emphasize that most black-hole masses have been measured using X-ray novae, for which $q > 1$, in the quiescent state. In this case, the effect of X-ray heating is small, and the error in q affects the value of m_x when $q > 1$ only weakly [see (2)]. As a result, the influence of systematic errors on q values derived from the rotational broadening of absorption lines is usually insignificant.

The orbital inclination i is most affected by systematic errors. A method for determining i from the optical light curves of X-ray binaries, whose shapes are determined primarily by the ellipticity of the optical star, was proposed in [18, 19], and is now being widely used to derive the masses of black holes in binary systems [1, 11]. The main source of systematic errors in the i values derived using this method is the contribution of gaseous structures (such as the accretion disk, gas streams, and the region of interaction between the stream and disk) to the total optical or infrared luminosity of the system. This contribution can be estimated spectrophotometrically by comparing the equivalent widths of absorption lines in the spectrum of the binary system with the corresponding equivalent widths in the spectrum of an isolated star of the same spectral and luminosity class. Unfortunately, the contribution of such gaseous structures can exceed 50% in the case of X-ray novae—binaries with low-mass cool stars [20], and the orbital variability of the emission of these structures is complex [21]. As a result, the systematic errors in i, and consequently in the mass m_x, become considerable. For example, the mass of the black hole m_x in the X-ray nova GRO J0422+32, estimated using two different methods, varies from $2.5–4M_\odot$ to $> 9M_\odot$ [20]. In the case of quasi-stationary close X-ray binaries with massive, hot stars (Cyg X-1, LMC X-1, and LMC X-3), the contribution of the optical radiation of gaseous structures is small ($< 2\%$), but the optical light curves of these systems suffer from the effects of absorption of the light from the optical star by the gaseous structures [22], which also leads to systematic errors in the orbital inclinations of these close binaries. In addition, the optical stars in the Cyg X-1 and LMC X-1 systems do not quite entirely fill their Roche lobes. This also introduces an extra systematic error into i, so that information about the distance to the system is required [11].

The new method for determining q and i presented in [16, 17], based on analysis of the orbital variability of absorption-line profiles in the spectrum of the optical star in a close X-ray binary, does not depend on the contribution of gaseous structures to the total luminosity of the system. Therefore, high-quality spectroscopy of close X-ray binaries with high spectral resolution $R = 50\,000–100\,000$ using the largest new-generation telescopes should enable us to appreciably reduce the effect of systematic errors and obtain the most trustworthy estimates of black-hole masses in close binary systems.

Another opportunity for independently determining the orbital inclinations i of X-ray binaries is the use of new, more accurate information on their distances d, which will be provided by next-generation astrometric space observatories (SIMA, GAIA, etc.). Knowledge of the distance d, interstellar absorption A_v, apparent magnitude m_v, and the contribution of gaseous structures to the system's luminosity enables determination of the average radius of the optical star R_v. This gives us a relation between q, μ, and i [11, 23] (where μ is the degree of filling of the Roche lobe by the optical star):

$$\sin i = \left(\frac{0.38\mu}{R_v}\right)\left(\frac{GP^2 f_v(m)}{4\pi^2}\right)^{1/3}\left(\frac{1+q}{q^{1.208}}\right). \quad (4)$$

Since we can assume for X-ray novae that $\mu = 1$ [2] and the value of q can be determined independently from the rotational broadening of absorption lines in the optical spectrum [see (3)], Eq. (4) can be used to obtain an independent determination of the orbital inclination i.

Thus, the probable presence of considerable systematic errors in dynamic black-hole masses currently prevents us from firmly establishing the black-hole mass distribution. We shall consider below only two limiting cases of the distribution of black-hole masses in close binaries: (1) a sharp δ-function-like peak near some specified value $\sim 9–10M_\odot$ and (2) a uniform distribution over a broad range ($4–15M_\odot$).

3. DIFFERENCES IN THE OBSERVABLE MANIFESTATIONS OF ACCRETING NEUTRON STARS AND BLACK HOLES

Since it will be important for us that the masses of black holes and of neutron stars have different lower limits, it is appropriate to underline here the fundamental differences in the observable manifestations of these two types of relativistic objects. As was noted above, the masses of approximately 40 compact objects in binary systems—18 black holes and more than 20 neutron stars—have currently been

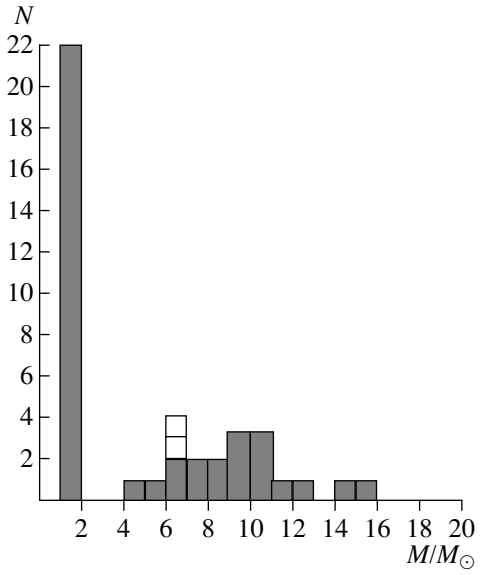

Fig. 1. Observed mass distribution of relativistic objects in close binary systems. The neutron stars are concentrated in the narrow interval $M = (1–2)M_\odot$, while black holes in close binaries are found in the interval $4–16M_\odot$. The masses of isolated black holes derived from microlensing observations are marked by hollow squares.

measured. It is a remarkable fact that the observed features of the accreting neutron stars and black holes differ, in accordance with the quantitative predictions of Einsteinian general relativity: there is a gap in the observed manifestations of compact relativistic objects near a mass of $3M_\odot$ (the absolute upper limit for the mass of a neutron star following from general relativity). In all cases when the mass of an X-ray pulsar, type I X-ray burster, or radio pulsar (phenomena showing clear signs that we are observing the surface of a relativistic compact object) is determined, the corresponding masses do not exceed $3M_\odot$, in good agreement with general relativity. The large number of known neutron-star masses (over 20) makes this result statistically significant. On the other hand, none of the 18 massive ($> 3M_\odot$) relativistic objects (black-hole candidates) is associated with an X-ray pulsar, type I X-ray burster, or radio pulsar. Therefore, none of the black-hole candidates shows features associated with an observable surface, as should be the case for black holes in general relativity. The increasing number of such objects (currently 18) confirms the high statistical significance of this result. Of course, the absence of clear features associated with an observable surface represents a necessary, but not sufficient, indication of a black hole.

We emphasize that there are also finer observational spectral differences between accreting neutron stars and black holes, as well as differences in the rapid variability of their X-ray emission (see, for example, [24, 25]). These differences are also consistent with the idea that neutron stars have observable surfaces, while black holes do not.

Therefore, all the necessary conditions imposed by general relativity on the observable manifestations of accreting neutron stars and black holes are satisfied. This strengthens our certainty in the existence of black holes in nature. Further, there is hope that sufficient criteria for the observational identification of black holes will be obtained very soon using the X-ray space interferometer [26] and via the detection of bursts of gravitational radiation due to merging black holes in binaries using gravitational-wave interferometers such as LIGO, VIRGO, and LISA (for more details, see the review [27]).

4. THE OBSERVED BLACK-HOLE MASS FUNCTION

Thus, modern astronomical data provide a basis for a discussion of the observed mass function for stellar-mass black holes (Fig. 1). The apparent mass distribution is in the range ~ 4 to $\sim 20M_\odot$, with no significant concentration at any particular mass. Since the systematic errors in the mass of an invisible companion in a close binary (especially ambiguity in the orbital inclination) can appreciably distort the true distribution, we shall consider two limiting cases: (a) a narrow distribution of masses around some specified value M_0 (for definiteness, we adopt $M_0 = 10M_\odot$) and (b) a uniform distribution over some range $M_{\min} - M_{\max}$, where $M_{\min} = (3–4)M_\odot$ and $M_{\max} = (15–20)M_\odot$.

Case (a): a narrow mass function around a specified value M_0, $dN/dM \sim \delta(M - M_0)$. There is no fundamental physical justification for the realization of such a distribution. Moreover, the observed X-ray luminosity function of massive close binary systems in other galaxies is clearly inconsistent with this hypothesis (see below).

Case (b): a flat (or almost flat) mass function $dN/dM \sim M^{-\beta}$, where $\beta \approx 0$, $M_{\min} < M < M_{\max}$. Such a distribution seems more probable, since the masses of collapsing supernova progenitors are distributed over some interval, and the fraction of the stellar mass that collapses into the black hole can depend substantially on the physical conditions of the collapse (such as rotation, magnetic field, etc.). Accordingly, we will consider this case to be realized below.

4.1. Initial Black-Hole Mass Function: Direct Calculations

The mass function is a fundamental characteristic of black holes. Modern theoretical concepts about the collapse of stellar cores are incomplete, and cannot predict unambiguously the masses of the resulting compact remnants. For example, a bimodal initial mass function (IMF) for the compact remnants (with peaks at $M_{NS} = 1.28M_\odot$ for neutron stars and $M_{BH} = 1.78M_\odot$ for black holes) was obtained in calculations of collapsing type II supernovae [28], in clear contradiction with the absence of observed black-hole candidates with masses below 3–$4M_\odot$. In contrast, based on certain assumptions about the masses of the resulting black holes, Fryer and Kalogera [10] obtained a broad, continuous distribution of black-hole masses up to 10–$15M_\odot$, without any deficit of objects with masses 1.5–$3M_\odot$. Neither of these theoretical distributions are in satisfactory agreement with the observations. The origin of these discrepancies may lie in observational selection effects. For example, Fryer and Kalogera [10] suppose that, if a black hole acquires some additional velocity ("kick") during its formation, low-mass black holes in binary systems will have a lower probability of surviving after the ejection of the massive envelope. This argument is doubtful, since this effect should be even stronger during the formation of neutron stars in binary systems [5], in evident contradiction with the observed pattern.

A possible physical explanation for the absence of observed masses of compact objects in the range 1.5–$3M_\odot$ was proposed in [9], which considers a magnetorotational mechanism for supernova explosions [29] and a fairly soft equation of state for neutron stars, with a limiting mass of $M_{\max} \approx 1.6M_\odot$.

However, there is no doubt that all such calculations are model-dependent, and, moreover, do not adequately take into account the effects of rotation, the magnetic field, the possible accretion of matter from the ejected envelope, and so on. It is likely that the derivation of the black-hole IMF will require the use of phenomenological data on the core masses and other physical characteristics of supernova progenitors derived from observations [5].

Nevertheless, it seems useful to analyze various hypotheses about the black-hole IMF and compare the results with observations. It is well known that the stellar IMF has the power-law form $f(M)_i = (dN/dM)_i \propto M^{-\alpha_i}$. The Salpeter IMF has an index (slope of the differential mass function) $\alpha_i = 2.35$ for stars with masses up to $10M_\odot$ in the solar neighborhood, and is in agreement with modern observations. The slope of the IMF for more massive stars becomes steeper. (This should probably be treated only as a trend due to the large errors in the masses of early-type stars and insufficient statistics.) For example, the Miller–Scalo IMF yields $\alpha_i = 2.5$ for stars with $M \sim 10M_\odot$. Some astronomers (e.g. B. Elmegreen, *et al.*) believe that the stellar IMF is a manifestation of the universal character of star formation in the turbulent self-gravitating interstellar medium in galaxies (see the recent review [30] and references therein). In addition, it is well known that the stellar winds of massive OB stars carry away a substantial fraction of the initial mass of the main-sequence stars, and the masses of supernova-progenitor cores are distributed over a wide range [5]. It is easy to see that power-law dependences for the fundamental parameters of stars (such as their luminosity and radius) on their mass can lead to a stellar mass distribution at the end of their thermonuclear evolution (just before the collapse) that also has a power-law form. Therefore, the power-law form of the black-hole IMF is admissible theoretically, but does not follow from any general physical arguments.

4.2. Variations in the Black-Hole Mass in the Course of Subsequent Evolution

The mass of a black hole that has formed in any way can either (a) increase due to accretion of matter (or, more precisely, energy) onto the black hole or (b) decrease due to quantum evaporation [31]. The mass M of an isolated black hole moving with speed v through an interstellar medium with density ρ and sound speed v_s increasess due to Bondi–Hoyle accretion $\dot{M}^+ \propto \rho v M^2/(v_s^2 + v^2)^2$. In the typical case, $v_s < 1$ km/s, and the velocity dispersion of massive stars in the Galactic disk (which can give birth to black holes at the end of their evolution) is of the order of 10 km/s, so that $\dot{M}^+ \sim \rho M^2/v^3 \sim 10^{13}$g/s $= 10^{-13}M_\odot$/yr for an isolated black hole with a mass of a few solar masses moving through a medium with a characteristic density of 10^{-23} g/cm^3. Therefore, the increase in the masses of isolated black holes in the Galaxy can be neglected. A unique opportunity to measure the masses of isolated black holes via observations of gravitational microlensing has recently appeared [32, 33]. This method has enabled the measurement of the masses of two black-hole candidates using the microlensing events MACHO-96-BLG-5 ($M = 6^{+10}_{-3}M_\odot$) and MACHO-98-BLG-6 ($M = 6^{+7}_{-3}M_\odot$) [32]. The corresponding values are also plotted in Fig. 1 (hollow squares).

The mass of a black hole in a close binary system can increase due to the accretion of matter from its companion. In the case of low-mass binaries containing black holes (such as X-ray novae), the average accretion rate is determined by the evolution of the

binary orbit as its orbital angular momentum is carried away by gravitational radiation or the magnetized stellar wind of the optical star, and will be of the order of $\dot{M}^+ \sim (10^{-9}$–$10^{-10})\ M_\odot$/yr. The accretion rate can be greater for black holes in massive close binaries (such as Cyg X-1 and SS 433). If the standard regime of accretion in a thin disk is realized [34, 35], the rate of increase of the black-hole mass will be limited by the Eddington luminosity (about $10^{-7}\ M_\odot$/yr for $M = 10M_\odot$). In the case of advection-dominated accretion, the rate of increase in the mass can be even greater. However, evolutionary considerations indicate that there should be substantially fewer accreting black holes in massive close binary systems in our Galaxy [36]. The increase in the mass of a black hole will obviously be determined by the duration of the accretion stage ($\sim (10^7$–$10^8)$ yr for low-mass and $\sim 10^5$ yr for massive close binaries). Therefore, in the case of black holes in close binaries, in first approximation we can neglect the possible increase in their mass by about 10%.

4.3. The Black-Hole Mass Function and the Luminosity Function of X-Ray Sources in Galaxies

The high angular resolution of the modern CHANDRA and XMM X-ray telescopes makes it possible to study individual X-ray sources in other galaxies and, in particular, to construct their distribution over the observed X-ray luminosity; see, for example, [37], as well as [38], which presents the X-ray luminosity function constructed using the HRI instrument onboard the ROSAT satellite. These and other works (see also, for example, the recent review [39]) have shown that the luminosity function of point-like X-ray sources in various galaxies has a power-law form $dN/dL_x \propto L_x^{-\beta}$ over a wide range of luminosities 10^{36} to $\sim 10^{40}$ erg/s, with the index being $\beta \sim 1.5$–1.7. The hypothesis that there exists a universal X-ray luminosity function with index $\beta \approx 1.6$ for the population of binary systems in galaxies was put forward and argued in [37]. As was shown in [40], the existence of a universal power-law for this X-ray luminosity function can be explained by the nature of accretion onto compact objects in massive close binary systems. The characteristic properties of the observed X-ray luminosity function are (1) the absence of a visible break at $L_x \approx 10^{38}$ erg/s (the Eddington limit for accretion onto a neutron star) and (2) a sharp cutoff in the function at a luminosity of $\sim(2 \times 10^{39} - 2 \times 10^{40})$ erg/s. Although the corresponding observations may be statistically incomplete, let us consider what we can deduce about the masses of accreting black holes in binary systems based on the X-ray luminosity function.

Let us begin with the cutoff of the observed luminosity at $\sim 2 \times 10^{40}$ erg/s. Let us suppose that this maximum luminosity is equal to the Eddington luminosity, $L_{\rm Edd} \approx 10^{38}(M/M_\odot)$ erg/s. Depending on its inclination, the luminosity of a standard accretion disk can be a factor of three to six higher than the nominal Eddington luminosity (see discussion in [37]). The maximum mass of the corresponding black holes would then be $M_{\max} \sim (20$–$30)M_\odot$. We believe that the observation of such bright X-ray sources in many galaxies is difficult to reconcile with the hypothesis that the black-hole masses are concentrated near the value $\sim 10M_\odot$, suggesting a fairly broad distribution of black-hole masses is more likely. An alternative explanation for the ultrabright X-ray sources observed in other galaxies is that they are microquasars whose jets are directed toward the observer (see the discussion in the review [41] and references therein). In this case, the true X-ray luminosity of the source should be a factor of at least $1 - \cos\theta$ lower than the luminosity derived from the received radiation flux assuming spherical symmetry of the source (where θ is the opening angle of the collimation cone of the radiation). The estimates of [42] show that this hypothesis requires unreasonably broad collimation of the radiation $\theta \sim 30^\circ$–60° in order to obtain agreement with the statistics of the observed ultrabright X-ray sources. In addition, the microquasar hypothesis is not consistent (at present) with the observed absence of a break in the X-ray luminosity function near $\sim 10^{38}$ erg/s.

The analysis of the X-ray luminosity function presented in Fig. 5 of [37] shows that the absolute value of the index characterizing the slope of the function dN/dL_x becomes greater than the mean value -1.6 at a luminosity of $\sim 2 \times 10^{39}$ erg/s, namely, $dN/dL_x \propto L_x^{-2\ldots-2.2}$. The following two conclusions can be drawn from this fact. First, including the factor of three to six noted above when interpreting the observed luminosity of an accretion disk radiating at the Eddington limit, a luminosity of 2×10^{39} erg/s corresponds to a black-hole mass of 3–$4M_\odot$. Second, if we assume that all ultrabright X-ray sources with $L_x > 2 \times 10^{39}$ erg/s are actually close binary systems with black holes whose luminosities are about equal to the Eddington luminosity, then $dN/dL_x \propto dN/dM$, and the observed slope of the X-ray luminosity function at high luminosities should directly reflect the distribution of black-hole masses in close binaries: $dN/dM \sim M^{-2\ldots-2.2}$. Since the increase in the masses of black holes in massive close binaries (which seem to correspond to ultrabright X-ray sources) is small during the accretion stage, the corresponding distribution should reflect the initial form of the black-hole mass function: $f_0(M) \propto M^{-2\ldots-2.2}$.

Finally, if accretion onto a black hole in a close binary occurs in a subcritical regime, the X-ray luminosity for standard disk accretion is simply $L_x = \eta \dot{M} c^2$, where the coefficient of proportionality depends on the rotation of the black-hole ($\eta \approx 0.06$ for nonrotating and 0.42 for maximally rotating black holes). In this case, the luminosity function of such X-ray sources does not depend on the mass of the black hole and is determined, as in the case of neutron stars, by the mass distribution of the optical components in the binaries and the dependence of the accretion rate on these masses [40]. This makes the absence of a break at the value of $L_{\rm Edd}$ for $1-2\ M_\odot$ quite natural. A cutoff in the luminosity function is expected at higher luminosities, determined by the limiting (Eddington) X-ray luminosity for the black hole with minimum mass. If $M_{\rm min} = 3-4 M_\odot$, the corresponding value could be a few times 10^{39} erg/s. The broad distribution of black-hole masses in close X-ray binaries is a supplementary factor that smooths the sharp break.

Therefore, we have arrived at two important conclusions: the distribution of black-hole masses in binary systems, $dN/dM \sim M^{-2...-2.2}$, derived from the observed X-ray luminosity function for ultrabright X-ray sources with $L_x > 2 \times 10^{39}$ erg/s in other galaxies (a) is consistent with the black-hole mass range $4-20\ M_\odot$ obtained from dynamical measurements and (b) is not consistent with a uniform distribution for the dynamical masses of black holes in close binaries within this same range. This latter conclusion can be explained by various selective effects, such as the possibility that the evolution of massive (ultrabright X-ray sources) and low-mass (most close X-ray binaries with known masses for their black-hole candidates) X-ray binaries proceeds along different paths. However, we can also seek a physical origin for the observed discrepancy that is not related to evolutionary processes. With this aim in view, let us consider the hypothesis of enhanced evaporation of stellar-mass black holes.

5. ENHANCED EVAPORATION OF BLACK HOLES IN CERTAIN MODERN MODELS OF GRAVITY

In the framework of a classical, four-dimensional, Einsteinian theory of gravity, the quantum evaporation of stellar-mass black holes is negligible, since the characteristic time for Hawking evaporation, which is of the order of $\tau \sim t_{Pl}(M/m_{Pl})^3$ (where $t_{Pl} \sim 10^{-43}$ s and $m_{Pl} \sim 10^{-5}$ g are the fundamental Planck time and mass), becomes shorter than the current age of the Universe $t_H \sim 14 \times 10^9$ yr only for objects with masses below $\sim 10^{15}$ g (a detailed consideration of black-hole evaporation in the framework of general relativity can be found, for example, in [43]). Consequently, if we neglect quantum evaporation, the observed spectrum of black-hole masses in close binary systems should reflect the initial mass function of black holes in these systems. From this point of view, the observed flat spectrum over a broad range of masses leads us to the conclusion [5] that not only the mass of the supernova progenitor but also a number of other physical parameters (rotation, magnetic field, etc.) determine the mass of the black hole formed during the collapse of the stellar core.

Modern attempts to devise a unified theory of physical interactions have primarily promoted superstring theory as the most promising possibility (see the review [44]). This is considered to be the most realistic version of a quantum theory of gravity (which must describe, in particular, the evaporation of black holes). The concepts of superstring theory, always formulated in a multi-dimensional space, has led recently to multi-dimensional models of gravity with a macroscopic additional dimension (see the review [45]). Roughly speaking, these models can be subdivided into two broad classes: models with a factorized geometry (of the ADD type [46]) and models with a nonfactorized geometry (of the RS type [47]). The latter are preferable from the viewpoint of modern cosmology [45], and we shall accordingly consider black holes within the RS approach. In the simplest versions of this model, the observable physical world (i.e., particles and fields apart from gravity) is localized on a four-dimensional surface (the so-called brane) imbedded into an extra dimension (the so-called bulk) whose geometry is described by an anti-deSitter (AdS) metric. The four-dimensional metric described by classical general relativity is induced on the RS brane. The characteristic scale of the additional dimension (warp factor) is just the inverse of the radius of curvature $\mathcal{L}$ of the five-dimensional AdS metric. An extremely important (and, probably, the most fundamental) property discussed in recent years is the correspondence between supergravity in a five-dimensional AdS space and conformal field theory (supersymmetric Yang–Mills theory) on a four-brane (the so-called AdS/CFT correspondence; for more details, see the review [48] and references therein).

Attempts to derive static black-hole-type solutions within RS models have thus far been unsuccessful: "black cigar" type solutions (a black hole on a four-brane that asymptotically transforms into the AdS space [49]) are unstable [50], and clearly unable to describe the result of the collapse of a massive stellar core on the brane. There have been some attempts to obtain numerical solutions for black holes localized on a four-dimensional RS brane, but hints of a static solution were obtained numerically only

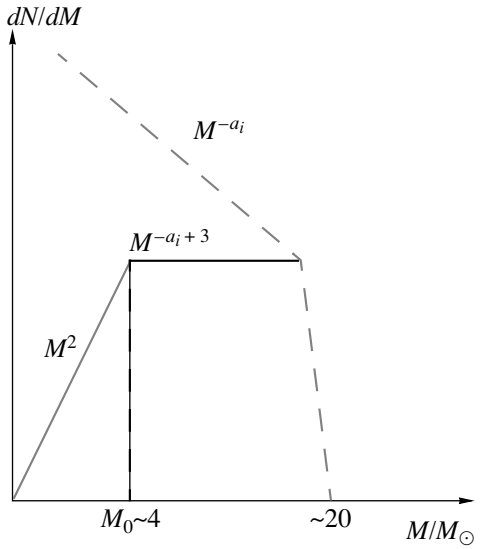

Fig. 2. Qualitative shape of the expected stationary black-hole mass distribution (solid curve) with a power-law initial form $(dN/dM)_i \sim M^{-a_i}$ (dashed curve) in a model with enhanced evaporation of the black hole on the RS2 brane. The mass M_0 corresponds to the minimum mass of a black hole that can evaporate over the Hubble time.

for black holes whose horizons were less than $\mathcal{L}$ in size [51].

Analysis of the classical evaporation of black holes on the RS brane [53] shows that, if the AdS/CFT correspondence is valid for black holes, astrophysically interesting features of their evaporation will appear. Namely, the classical evaporation of four-dimensional black holes can occur much more rapidly (at least, as long as the radius of the horizon of the four-dimensional black hole on the brane is greater than $\mathcal{L}$). The evaporation time in this model is

$$\tau \equiv \frac{M}{\dot{M}} \sim \frac{1}{\mathcal{L}^2}(G_4 M)^3 \sim 1\mathrm{yr}\left(\frac{M}{M_\odot}\right)^3\left(\frac{1\mathrm{mm}}{\mathcal{L}}\right)^2, \quad (5)$$

where G_4 is the effective Newtonian gravitational constant on the brane. An independent field-theoretical analysis [53], also based on the AdS/CFT correspondence, resulted in a qualitatively similar expression for the evaporation time:

$$\tau \simeq 10^2\mathrm{yr}\left(\frac{M}{M_\odot}\right)^3\left(\frac{1\mathrm{mm}}{\mathcal{L}}\right)^2. \quad (6)$$

The physical reason for the increase in the rate of evaporation of black holes in these models is that the evaporation rate increases in proportion to the number of degrees of freedom of the corresponding four-dimensional conformal field theory on the brane: $\propto (\mathcal{L}/l_{Pl})^2$, where $l_{Pl} \approx 10^{-33}$ cm is the classical Planck length. The discrepancy in the coefficients in the above formulas is due to the model allowance for the number of degrees of freedom in [53].

Note that the evaporation of a black hole into CFT modes produces essentially low-energy Kaluza–Klein gravitons, which are weakly coupled to the fields of ordinary matter, and are therefore unobservable by direct astrophysical methods. Moreover, the accelerated evaporation of black holes no longer takes place when the radius of the causality horizon approaches the size $\mathcal{L}$. These interesting problems are currently poorly understood (see, for example, the paper [54], in which results different from those of [53] were obtained).

If the application of the AdS/CFT correspondence to black holes is justified and the corresponding hypotheses are valid, the existence of stellar-mass black holes itself imposes extremely strong constraints on the value of the fundamental AdS radius, namely, $\mathcal{L} < 10^{-3} - 10^{-4}$ mm (for the model of [53]), while the modern laboratory constraints are $\mathcal{L} \lesssim 0.1$ mm [55].

6. INITIAL BLACK-HOLE MASS FUNCTION: THE INVERSE PROBLEM

In spite of the hypothetical nature of the above concepts (starting from the adequacy of describing the Universe using models with macroscopic additional dimensions!), let us try to use them to explain the observed spectrum of the dynamically-measured masses of black holes in close binary systems. Namely, *let us suppose that the observed absence of black holes with masses below $4M_\odot$ is due to their rapid evaporation in the RS model*. Consequently, black holes with smaller masses cannot be observed, at least in old close binary systems. Of course, the collapse of a massive stellar core at the end of its evolution can give birth to a black hole with an even lower mass, but its lifetime will be short due to the enhanced evaporation. We emphasize again that the evaporation of black holes in this model occurs in unobservable CFT modes; i.e., from the viewpoint of a distant observer, the mass of the black hole decreases without any other detectable effects. The contribution of the possible evaporation of stellar-mass black holes to the total energy budget of the Galaxy is also negligible. Let us adopt the extreme assumption that all black holes formed via the evolution of ordinary stars over Hubble time have evaporated. For our estimates, we take the average rate of star formation from baryons in the Galaxy to be $\sim 1M_\odot$/yr and the lower limit for the initial stellar mass that can give birth to a black hole at the end of its evolution to be $30M_\odot$. Then, for a Salpeter initial mass function, the mass of baryons transformed into black holes over

Hubble time should be about 1% of the total baryonic mass of the Galaxy.

Let us estimate the initial black-hole mass function $f_0(M)$ that is required to satisfy the observed black-hole mass distribution $f(M) = dN/dM \approx$ const for a given mass-variation law dM/dt. In the stationary case, the evolution of the one-dimensional distribution function is described by the kinetic equation

$$\frac{\partial}{\partial M}\left[f(M)\dot{M}\right] = f_0(M), \tag{7}$$

which, in the case $\dot{M} < 0$ (evaporation), reduces to

$$f(M) = \frac{\int_M^{M_{\max}} f_0(M')dM'}{\dot{M}}, \quad M > M_{\min}. \tag{8}$$

When $M \le M_{\min}$, the form of the stationary distribution does not depend on the initial mass function, and is determined only by the black-hole mass-variation law:

$$f(M) = \frac{\int_{M_{\min}}^{M_{\max}} f_0(M')dM'}{\dot{M}} = \frac{\text{const}}{\dot{M}}, \quad M \le M_{\min}. \tag{9}$$

If the rate of evaporation is higher than the rate of increase in the mass (in close binary systems with an average accretion rate of $\sim 10^{-10}M_\odot$/yr, this condition is satisfied when $\mathcal{L} \gtrsim 10^{-2}$ mm for the model of [53]), then $dM/dt = \dot{M}^- \propto M^{-2}$. Assuming a power-law form for the initial black-hole mass function $f_0(M) \propto M^{-\alpha_i}$, we obtain $f(M) \sim M^{-\alpha_i+3}$ at $M > M_{\min}$ and $f(M) \sim M^2$ when $M \le M_{\min}$, as is illustrated in Fig. 2 (we assumed $M \ll M_{\max}$ in the above estimates). It is interesting that a flat distribution is obtained when the coefficient of the slope of the initial black-hole mass function is $\alpha_i \sim 3$, close in absolute value to the slopes of the initial mass function of main-sequence stars ($\alpha_i \approx -2.5$) and the mass function of black holes in massive close binaries derived from observations of ultrabright X-ray sources in other galaxies ($\alpha_i \approx -2 \ldots -2.2$).

The self-consistency of the hypothesis being considered can be tested as follows. The condition for the evaporation of a black hole with a mass below M_0 over the Hubble time in the model of [53] leads to a constraint on the AdS radius:

$$\left(\frac{\mathcal{L}}{1[\text{mm}]}\right)^2 \gtrsim 10^{-8}\left(\frac{M_0}{M_\odot}\right)^3. \tag{10}$$

Consequently, canceling out the factor $\mathcal{L}^2$ in the expression for the evaporation rate, we obtain

$$\left(\frac{dM}{dt}\right)^- \gtrsim 3\times 10^{-11}M_\odot/\text{yr}\left(\frac{M_0}{M_\odot}\right). \tag{11}$$

Therefore, the conditions for evaporation over Hubble time and for an excess of the evaporation rate over the average accretion rate in close binary systems are satisfied simultaneously when $M_0 \gtrsim 10M_\odot$. The value $M_0 \sim 4M_\odot$ adopted in our analysis is lower than but fairly close to this limit (given the considerable model uncertainty in the numerical coefficients in the formulas describing evaporation). On the other hand, fixing the value $M_0 = 4M_\odot$ yields $\mathcal{L} \gtrsim 5\times 10^{-4}$ mm and $\dot{M}^- \gtrsim 10^{-10}M_\odot$/yr, which are also consistent with the available constraints on $\mathcal{L}$ and the hypothesized decrease in the masses of black holes in close binaries due to their evaporation.

7. CONCLUSION

Analysis of the observed distribution of masses of relativistic objects (neutron stars and black holes) in close binary systems leads to the conclusion that the masses of neutron stars and black holes are distributed according to substantially different laws. The neutron-star masses are concentrated within the narrow range $1-2\ M_\odot$, while the black-hole masses are spread over the broad interval $4-15M_\odot$, without a concentration near any specific mass. The uncertainties in the dynamical masses of black holes are due primarily to systematic errors introduced by the methods used to estimate the orbital inclinations and the component-mass ratios of the close binaries, derived from the light curves and spectra of the optical stars (related to the model dependence of the contributions of gaseous structures to the total optical luminosity of the systems). These uncertainties can be reduced by using refined models for the orbital variability of the absorption-line profiles in the spectra of the optical stars [12, 16, 17] and by using high-resolution spectra ($R = 50\,000-100\,000$) obtained with large modern telescopes when comparing the modeled and observed profiles. More accurate distances to X-ray binaries measured by next-generation space astrometric observatories (such as SIMA, GAIA, etc.) will also facilitate determinations of the orbital inclinations i of the binary systems. It will also be useful to accumulate information about the masses of isolated black holes via observations of gravitational microlensing events.

The mass function of black holes in massive close binaries can also be derived from observations of ultrabright ($L_x > 2\times 10^{39}$ erg/s) X-ray sources in other galaxies [37]. Assuming that these sources represent massive X-ray binaries radiating at the Eddington luminosity, the observed slope of the luminosity function at luminosities of $2\times 10^{39}-2\times 10^{40}$ erg/s leads to a power-law black-hole mass function $dN/dM \sim M^{-2.2}$, in contradiction with the lack of concentration of the black-hole masses near

the lower limit $4M_\odot$, as follows from the derived dynamical masses of black holes in close binaries (most of which are low-mass systems).

The characteristic features of the black-hole mass distributions in these two cases can be reconciled under the hypothesis [52, 53] that the evaporation of black holes is enhanced on the RS2 brane due to the large number of (unobservable) CFT modes that appear in the extrapolation of AdS/CFT correspondence to black holes on the brane. This model can also explain the absence of observed black holes with masses $< 4M_\odot$ in low-mass close binaries with low average accretion rates. This hypothesis can be verified by searching for black holes with lower masses (both isolated and in binary systems).

Thus, the reliable determination of the mass function of compact relativistic objects in close binary systems is a very important observational problem of modern astrophysics. This function can be used both to test the general relativistic theory of the formation of neutron stars and black holes during the collapse of the cores of massive stars and to verify theories of gravity that are fundamental in a deeper sense.

ACKNOWLEDGMENTS

We are grateful to A.A. Starobinsky and D.V. Galtsov for fruitful discussions and comments. KAP was supported by the Russian Foundation for Basic Research (projects nos. 02-02-16500, 03-02-16110, and 03-02-16068). AMC was supported by the Russian Foundation for Basic Research (project no. 02-02-17524) and the Program of Support of Leading Scientific Schools of Russia (project 00-15-96-553).

REFERENCES

1. J. A. Orosz, *IAU Symp. 212: A Massive Star Odyssey, From Main Sequence to Supernova,* Ed. by K. A. van der Hucht, A. Herrero, and C. Esteban (Astronomical Society of the Pacific, San Francisco, 2003).
2. A. M. Cherepashchuk, Usp. Fiz. Nauk **173**, 345 (2003) [Phys. Usp. **46**, 335 (2003)].
3. C. D. Bailyn, R. K. Jain, P. Coppi, and J. A. Orosz, Astrophys. J. **499**, 367 (1998).
4. A. M. Cherepashchuk, *Proc. Intl. Conf. in Honor of Prof. A. G. Massevitch: Modern Problems of Stellar Evolution,* Ed. by D. S. Wiebe (Zvenigorod-Moscow, 1998), p. 198.
5. A. M. Cherepashchuk, Astron. Zh. **78**, 145 (2001) [Astron. Rep. **45**, 120 (2001)].
6. A. V. Tutukov and L. R. Yungel'son, Nauchn. Inform. Astron. Soveta Akad. Nauk SSSR **27**, 58 (1973).
7. E. P. J. van den Heuvel and J. Heise, Nature Phys. Sci. **239**, 67 (1972).
8. A. V. Tutukov and A. M. Cherepashchuk, Astron. Zh. **80**, 419 (2003) [Astron. Rep. **47**, 386 (2003)].
9. K. A. Postnov and M. E. Prokhorov, Astron. Zh. **78**, 1025 (2001) [Astron. Rep. **45**, 899 (2001)].
10. C. L. Fryer and V. Kalogera, Astrophys. J. **554**, 548 (2001).
11. A. M. Cherepashchuk, Usp. Fiz. Nauk **166**, 809 (1996) [Phys. Usp. **39**, 759 (1996)].
12. E. A. Antokhina and A. M. Cherepashchuk, Astron. Zh. **71**, 420 (1994) [Astron. Rep. **38**, 367 (1994)].
13. E. A. Antokhina, A. M. Cherepashchuk, and V. Shimanskiĭ, Izv. Ross. Akad. Nauk, Ser. Fiz. **67**, 293 (2003).
14. M. Milgrom, Astron. Astrophys. **70**, 763 (1978).
15. M. H. Van Kerkwijk, J. Van Paradijs, E. J. Zuiderwijk, *et al.*, Astron. Astrophys. **303**, 483 (1995).
16. E. A. Antokhina and A. M. Cherepashchuk, Pis'ma Astron. Zh. **23**, 889 (1997) [Astron. Lett. **23**, 773 (1997)].
17. T. Shahbaz, Mon. Not. R. Astron. Soc. **298**, 153 (1998).
18. V. M. Lyutyĭ, R. A. Syunyaev, and A. M. Cherepashchuk, Astron. Zh. **50**, 3 (1973) [Sov. Astron. **17**, 1 (1973)].
19. V. M. Lyutyĭ, R. A. Syunyaev, and A. M. Cherepashchuk, Astron. Zh. **51**, 1150 (1974) [Sov. Astron. **18**, 684 (1974)].
20. A. M. Cherepashchuk, Space Sci. Rev. **93**, 473 (2000).
21. T. S. Khruzina, A. M. Cherepashchuk, D. V. Bisikalo, *et al.*, Astron. Zh. **78**, 625 (2001) [Astron. Rep. **45**, 538 (2001)].
22. A. V. Goncharskiĭ, S. Yu. Romanov, and A. M. Cherepashchuk, *Finite-Parametric Inverse Problems in Astrophysics* (Mosk. Gos. Univ., Moscow, 1991) [in Russian].
23. B. Paczyński, Astron. Astrophys. **34**, 161 (1974).
24. R. A. Sunyaev, *Black Holes in Binaries and Galactic Nuclei,* Ed. by L. Kaper, E. P. J. van den Heuvel, and P. A. Woudt (Springer, Berlin, 2001), p. 113.
25. Y. Tanaka, *Black Holes in Binaries and Galactic Nuclei*, Ed. by L. Kaper, E. P. J. van den Heuvel, and P. A. Woudt (Springer, Berlin, 2001), p. 141.
26. N. White, Nature **407**, 146 (2000).
27. L. P. Grishchyuk, B. M. Lipunov, K. A. Postnov, *et al.*, Usp. Fiz. Nauk **171**, 3 (2001) [Phys. Usp. **44**, 1 (2001)].
28. F. X. Timmes, S. E. Woosley, and T. A. Weaver, Astrophys. J. **457**, 834 (1996).
29. G. S. Bisnovatyĭ-Kogan, Astron. Zh. **47**, 813 (1970) [Sov. Astron. **14**, 652 (1970)].
30. Yu. N. Efremov and A. D. Chernin, Usp. Fiz. Nauk **173**, 3 (2003).
31. S. W. Hawking, Nature **248**, 30 (1970).
32. D. P. Bennett, A. C. Becker, J. L. Quinn, *et al.*, Astrophys. J. **579**, 639 (2002).
33. S. Mao, M. C. Smith, and P. Wozniak, Mon. Not. R. Astron. Soc. **329**, 349 (2002).
34. N. I. Shakura, Astron. Zh. **49**, 921 (1972) [Sov. Astron. **16**, 756 (1972)].
35. N. I. Shakura and R. A. Sunyaev, Astron. Astrophys. **24**, 337 (1973).

36. V. M. Lipunov, K. A. Postnov, and M. E. Prokhorov, Astrophys. Space Phys. Rev. **9**, 1 (1996).
37. H.-J. Grimm, M. R. Gilfanov, and R. A. Sunyaev, Mon. Not. R. Astron. Soc. **339**, 793 (2003).
38. T. P. Roberts and R. S. Warwick, Mon. Not. R. Astron. Soc. **315**, 98 (2000).
39. L. K. Summers, I. R. Stevens, D. K. Strickland, and T. M. Heckman, Mon. Not. R. Astron. Soc. (2003, in press); astro-ph/0303251.
40. K. A. Postnov, Pis'ma Astron. Zh. **29**, 424 (2003).
41. S. N. Fabrika, Astrophys. Space Phys. Rev. (2003, in press).
42. S. N. Fabrika and A. Mesheryakov, *IAU Symp. 205: Galaxies and their Constituents at the Highest Angular Resolutions,* Ed. by R. T. Schilizzi, S. N. Vogel, F. Paresce, and M. S. Elvis (Astronomical Society of the Pacific, San Francisco, 2001), p. 268.
43. I. D. Novikov and V. P. Frolov, *Physics of Black Holes* (Nauka, Moscow, 1988) [in Russian].
44. A. V. Marshakov, Usp. Fiz. Nauk **172**, 977 (2002).
45. V. A. Rubakov, Usp. Fiz. Nauk **171**, 913 (2001) [Phys. Usp. **44**, 871 (2001)].
46. N. Arkani-Hamed, S. Dimopoulos, and G. Dvali, Phys. Lett. B **429**, 263 (1998).
47. L. Randall and R. Sundrum, Phys. Rev. Lett. **83**, 4690 (1999).
48. E. T. Akhmedov, Usp. Fiz. Nauk **171**, 1005 (2001) [Phys. Usp. **44**, 955 (2001)].
49. A. Chamblin, S. W. Hawking, and H. S. Reall, Phys. Rev. D **61**, 065007 (2000).
50. G. T. Horowitz and K. Maeda, Phys. Rev. Lett. **87**, 131301 (2001).
51. H. Kudoh, T. Tanaka, and T. Nakamura, gr-qc/0301089 (2003).
52. T. Tanaka, Prog. Theor. Phys. Suppl. **148**, 307 (2003); gr-qc/0203082.
53. R. Emparan, J. García-Bellido, and N. Kaloper, J. High Energy Phys. **0301**, 079 (2003); hep-th/0212132.
54. R. Casadio, hep-th/0302171 (2003).
55. J. C. Long and J. C. Price, hep-ph/0303057 (2003).

Translated by Yu. Dumin